SALVADOR • v.30 • n.2 • JUL./DEZ. 2020 ISSN 0103 8117

BA&D

BAHIA ANÁLISE & DADOS

Big Data e Políticas Públicas

publicações SEI

BA&D

BAHIA ANÁLISE & DADOS

Bahia anál. dados | Salvador | v. 30 | n. 2 | p. 1-176 | jul.-dez. 2020

A *Bahia Análise & Dados* é uma publicação semestral da Superintendência de Estudos Econômicos e Sociais da Bahia (SEI), autarquia vinculada à Secretaria do Planejamento do Estado da Bahia. Todos os números podem ser visualizados no site da SEI (www.sei.ba.gov.br) no menu "Publicações". Os artigos publicados são de inteira responsabilidade de seus autores. As opiniões neles emitidas não exprimem, necessariamente, o ponto de vista da SEI. É permitida a reprodução total ou parcial dos textos desta revista, desde que a fonte original seja creditada de forma explícita. Esta publicação está indexada no *Library of Congress, Ulrich's International Periodicals Directory* e no sistema Qualis da Capes.

Bahia Análise & Dados, v. 1 (1991-)
Salvador: Superintendência de Estudos Econômicos e Sociais da Bahia, 2020.
v.30
n. 2
Semestral
ISSN 0103 8117

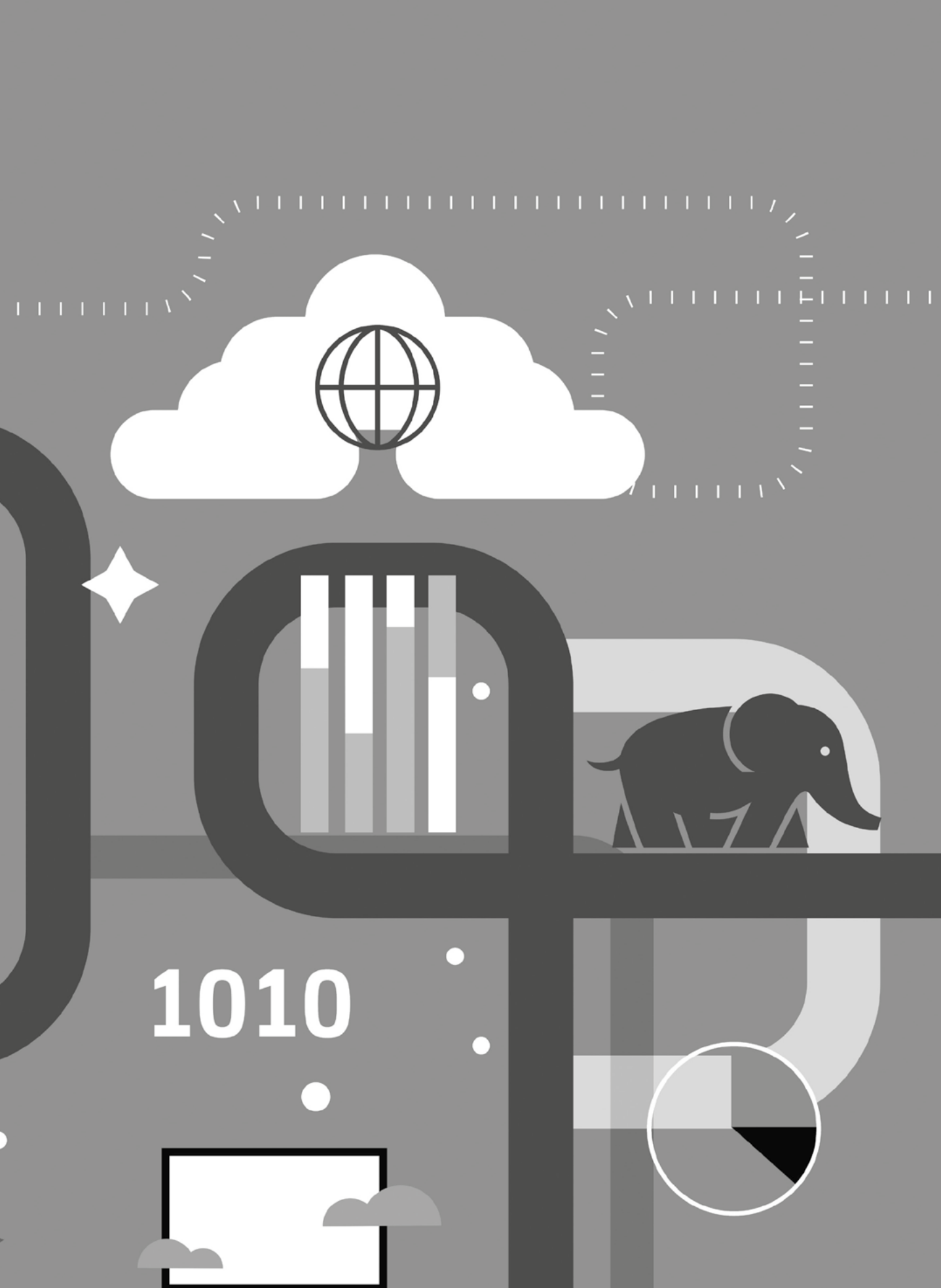
CDU 338 (813.8)

SEI 25 65
SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAIS DA BAHIA
25 ANOS DE SEI. 65 ANOS DE HISTÓRIA.

SECRETARIA DO PLANEJAMENTO

GOVERNO DO ESTADO
BAHIA. AQUI É TRABALHO.

Av. Luiz Viana Filho, 4ª Avenida, 435, CAB
Salvador (BA) Cep: 41.745-002
Tel.: (71) 3115 4822 Fax: (71) 3116 1781
www.sei.ba.gov.br sei@sei.ba.gov.br


1010

Apresentação

Prosseguindo em sua missão de estimular a produção e a publicação científica de conhecimento, a Superintendência de Estudos Econômicos e Sociais da Bahia (SEI) apresenta a nova edição da revista *Bahia Análise & Dados*, com o tema Big Data e Políticas Públicas, evocando as possibilidades de estudos, aplicações e análises a partir de grandes bancos de dados e sua incorporação à gestão pública.

Num período em que os reflexos globais da contaminação pelo novo coronavírus ainda afetam o mundo inteiro em diversos sentidos, sistemas de análise de grandes volumes de informações processadas e disponibilizadas por órgãos públicos e captadas nas redes sociais têm viabilizado, pelas possibilidades de se trabalhar com *big data*, estudos e ações governamentais para prevenção de novas infecções, por meio da identificação de determinados padrões ou modos de agir das pessoas.

Por diversas razões, inclusive por conta dos reflexos da covid-19 na sociedade global, o tema desta edição da *Bahia Análise & Dados* assume grande relevância e atualidade ao associar *big data* a políticas públicas e ao confirmar os aprendizados e autoaprendizados dos mecanismos de inteligência artificial em suas múltiplas possibilidades. Além disso, as questões aqui abordadas revelam as faces e os riscos dos processos de comunicação em massa influenciando comportamentos e determinando cenários incertos.

Dessa forma, a presente publicação traz, entre outros trabalhos, um estudo da Universidade da Integração Internacional da Lusofonia Afro-Brasileira (Unilab) e da Universidade Estadual do Ceará (UECE) objetivando conhecer, identificar e analisar padrões comportamentais da evolução de taxa de letalidade por covid-19 e de casos confirmados em municípios do Maciço de Baturité, no Ceará. Esta edição apresenta ainda uma entrevista com Wilson da Silva Gomes, professor doutor da Universidade Federal da Bahia (UFBA), sobre aspectos relacionados ao uso de inteligência artificial e de grandes volumes de dados e sua influência em comportamentos e decisões. A entrevista discute também se as redes sociais já estão comandando a vida das pessoas por meio de mensagens cifradas e direcionadas a grupos focais utilizando inteligência artificial e *big data*.

As questões propostas nesta edição também foram abordadas em artigos, que trataram de aspectos como a taxa de utilização municipal do Programa Bolsa Família na Bahia; a arquitetura para a integração de sistemas com fontes heterogêneas e *big data*; a utilização de técnicas de mosaico-síntese para analisar a violência no Brasil por meio dos dados de agressão resultantes em internações e óbitos; as aplicações no uso de *big data* pelos sistemas de saúde; e a caracterização da homogeneidade dos territórios de identidade da Bahia a partir de dados multivariados.

Assim, agradecemos a todos os que colaboraram com mais esta edição da revista *Bahia Análise & Dados* e reafirmamos o propósito da SEI de colaborar com a sociedade e com a gestão pública fornecendo dados e informações relevantes e que possam contribuir para uma melhor condução das políticas públicas voltadas ao bem-estar da população.

Big data e políticas públicas

ENÉZIO DE DEUS SILVA JÚNIOR

Doutor e mestre em Família na Sociedade Contemporânea, pela Universidade Católica do Salvador (UCSAL). Servidor efetivo da carreira de Especialista em Políticas Públicas e Gestão Governamental da Secretaria da Administração do Estado da Bahia (SAEB), em exercício na Superintendência de Estudos Econômicos e Sociais da Bahia (SEI).

GABRIEL OLIVEIRA BARBOSA

Mestre em Desenvolvimento Econômico, pela Universidade Federal do Paraná (UFPR) e graduado em Ciências Econômicas, pela Universidade Federal da Bahia (UFBA). Servidor efetivo da carreira de Especialista em Produção de Informações Econômicas, Sociais e Geoambientais da Superintendência de Estudos Econômicos e Sociais da Bahia (SEI).

PAULO JORGE CANAS RODRIGUES

Doutor em Estatística, pela Universidade Nova de Lisboa e mestre em Estatística, pela Universidade Técnica de Lisboa. Professor efetivo do Departamento de Estatística da Universidade Federal da Bahia (UFBA).

Coordenadores editoriais deste número da revista.

NO ÂMBITO organizacional privado, grandes volumes de informações vêm sendo utilizados, há muito, para pensar ou gerar novas e eficientes estratégias de consumo, por exemplo. Mais recentemente, seja quanto a políticas públicas e ações governamentais, seja nas interfaces ou parcerias público-privadas, começam a ser utilizadas ferramentas para estruturar grandes bancos de dados e gerar resultados em várias áreas, notadamente no campo da saúde, educação, justiça, economia, segurança pública, dentre outras.

Os termos Big Data e *Ciência de Dados*, que ficaram populares nos anos 2000 (VOLPATO; RUFINO; DIAS, 2014), representam a combinação de métodos estatísticos com algoritmos computacionais e têm como objetivo transformar grandes conjuntos de dados em informação e em tomada de decisão.

No decorrer dos anos, com a ampliação das pesquisas científicas sobre temas associados à *Ciência de Dados*, seus aspectos teórico-metodológicos, suas perspectivas de aplicabilidade, ferramentas/conceitos – como *machine learning*, *statistical learning, deep learning*, inteligência artificial –, técnicas avançadas de econometria, modelagem

Os dados são o ouro do futuro, visto que a economia digital estará pautada prioritariamente na análise de dados

estatística, entre outras, já vêm sendo investigadas e utilizadas para repensar e criar novas políticas públicas, a fim de que o seu alcance seja mais amplo e eficiente. Segundo Lima (2017, p. 79), "[...] no âmbito do governo, são muitas as possibilidades, uma vez que este retém grande parte dos dados da população. Os dados são o ouro do futuro, visto que a economia digital estará pautada prioritariamente na análise de dados". O que se deve buscar ao final é a geração de novos conhecimentos capazes de transformar a realidade social e melhorar a vida das pessoas.

Das experiências pioneiras e crescentes de utilização de técnicas relacionadas com Big Data para fins públicos percebe-se, a depender do fim almejado, a busca de padrões estatísticos com base em comportamentos diferenciados, reiterados ou previsíveis. Encontram-se também trabalhos que vão desde experiências de autocorrelação espacial até a realização de estudos de inteligência geoterritorial e/ou de análises interdisciplinares no campo social, com o cruzamento de informações de diferentes bancos de dados de instituições referenciadas. Temos observado instituições de pesquisa lançando mão dessas técnicas neste momento de pandemia pelo novo Coronavírus (SARSCoV2), buscando compreender problemas cujos reflexos se dão em escala global e cujas soluções ou possibilidades têm se revestido de grande complexidade, pois não há uma "fórmula mágica" de previsibilidade ou contenção de caráter geral.

Seja no âmbito da identificação de padrões comportamentais sob um enfoque regional (vez que o vírus não se dissemina da mesma forma em todos os lugares), seja no rastreamento/monitoramento dos casos entre municípios de uma mesma região para se tentar perceber determinada lógica processual (como evidencia o estudo dos pesquisadores Me. Erivando de Sena Ramos, Dr. John Hebert da Silva Felix, Dr. Francisco Horácio da Silva e Dra. Maria Helena de Paula, publicado nesta edição), a Inteligência Artificial e o Big Data vêm contribuindo para melhorar o cenário pandêmico e oferecer aos governos perspectivas de contenção da disseminação do SARSCoV2, embora a pandemia esteja revelando, a cada onda e em cada país, estado, região ou município, características complexas de compreensão e de atuação dos entes públicos em parceria com instituições de pesquisa, empresas e a sociedade civil.

Amoroso (2020, p. 1) exemplifica o esforço da ciência via Inteligência Artificial e Big Data, quanto a doenças infecciosas como a covid-19, com três possibilidades: a "visualização da saúde populacional" (expressão da autora), que pode contribuir na detecção, na tomada de decisões e na prevenção em tempo hábil; a esterilização de superfícies, realizada por robôs; o mapeamento genético e o uso de ferramentas para identificar proteínas virais que ajudem na descoberta de tratamentos e vacinas realmente eficazes. Segundo Amoroso (2020, p. 1),

> [...] cientistas estão utilizando a inteligência artificial (*IA*) e o *Big Data* para desenvolver formas de ajudar a identificar indivíduos infectados a limpar superfícies contaminadas e a tratar – assim como prevenir – a infecção. Algumas das ferramentas já estão sendo colocadas em prática; outras estão em desenvolvimento. A COVID-19 não é o último vírus que o mundo terá de combater e talvez a *IA* e o *Big Data* possam ajudar a evitar que uma pandemia como essa aconteça novamente.

O tema em questão, portanto, diante do cenário de pandemia que vivenciamos, reveste-se não somente de atualidade, mas de grande importância

No Brasil, a Fundação Oswaldo Cruz (FIOCRUZ) é uma das instituições de pesquisa que têm se destacado por suas contribuições para a compreensão e o reforço de estratégias de combate aos efeitos da pandemia, em permanente diálogo com a sociedade. A fundação participou, por exemplo, da Semana Nacional de Ciência e Tecnologia, (2020), que teve por tema "Inteligência Artificial: a nova fronteira da ciência brasileira". Na tarde do dia 21/10/2020, foi discutida, no evento, "A utilização da Inteligência Artificial e tomada de decisões sobre a pandemia no Brasil" e o vídeo de tal atividade pode ser acessado no site da instituição, assim como está disponível o Boletim Corona intitulado "Big Data, Inteligência de Dados e Pandemia", em formato também de vídeo, com o professor do Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE) e coordenador do Centro de Referência em Inteligência Empresarial (CRIE) da Universidade Federal do Rio de Janeiro (UFRJ) Dr. Marcos Cavalcanti.

O tema em questão, portanto, diante do cenário de pandemia que vivenciamos, reveste-se não somente de atualidade, mas de grande importância. Numa edição especial da Revista do Serviço Público da Escola Nacional de Administração Pública (ENAP), publicada este ano, pesquisadoras/es refletiram sobre como estão agindo ou devem se comportar os governos diante da pandemia da covid-19 e alguns trabalhos tangenciam o tema desta edição da *Bahia Análise & Dados*, como o artigo de Fabrício Martins Mendonça e Mário Antônio Ribeiro Dantas. Os autores, a partir da provocante indagação sobre onde estão a "Transformação Digital, Big Data, Inteligência Artificial e Análise de Dados" em face da covid-19, afirmam que a pandemia

> [...] requer muitos testes que poderiam ser minimizados através de monitoração de sinais vitais específicos, incluindo temperatura, frequência cardíaca, oxigenação e pressão sanguínea. Além disso, nossos experimentos mostraram que a adoção desses tópicos computacionais exige mudanças mais rápidas no comportamento digital, nos procedimentos de governos e pessoas, para serem bem-sucedidas como ambientes de aprimoramento e proteção da saúde individual (MENDONÇA; DIAS, 2020, p. 213).

O uso de grandes bancos de dados e a existência de instituições ou empresas que lidam com Big Data se diversificam a depender da área de

abrangência da política pública, do seu escopo ou do tipo de estudo que se pretende realizar. Na gestão pública, o seu uso remonta, segundo Ferrarezi e Tomacheski (2011, p. 3), à "[...] emergência, nos anos 1990, dos princípios da Nova Gestão Pública, propondo tornar as organizações públicas mais *accountables*, por meio de mecanismos de controle de resultados e uso de indicadores [...]", objetivando mensurar desempenho e gerar mais eficiência. Tomando a missão dos Tribunais de Contas do País como referência para refletir sobre a utilização do Big Data em suas estruturas, Castro (2018, p. 19) reconhece que as

> [...] organizações nacionais e internacionais começam a elaborar metodologias que possam avaliar o tamanho do Estado, o escopo de suas funções, a efetividade das políticas públicas, o nível de transparência, indicadores de governança e gestão, qualidade do gasto público, equilíbrio fiscal, dívida pública, Previdência e principalmente o controle social. O desenvolvimento concomitante da tecnologia da informação permite atualmente aos Tribunais de Contas a implementação de Big Data em suas estruturas, sendo que os órgãos são repositórios do maior número de informações da administração pública brasileira, referentes aos 5.570 municípios, 26 estados e o Distrito Federal.

Eight Data Science Index (2019), ao realizar um estudo panorâmico sobre a violência contra as mulheres em São Paulo, cruzou informações de bancos diferentes de dados para mapear regiões com maior incidência de comportamentos violentos com base no gênero feminino, horários com mais volume de crimes de tal natureza e possíveis padrões de vulnerabilidade detectados a partir dos achados e correlações. Segundo Bortolassi (2019, p. 1), quanto ao uso de bancos com grande volume de dados não estruturados na esfera pública, é possível

> [...] estabelecer correlações com base nisso, para poder compreender o problema, a causabilidade e, a partir daí, construir o seu projeto e a sua narrativa, de que forma você pretende abordar isso e implementar um modelo de política pública ou criar uma consciência na imprensa e na opinião pública a respeito do assunto.

A pandemia tem se mostrado um momento propício para fomentar tal diálogo ou mais abertura de comunicação sobre o assunto e, diante do tema proposto por esta edição da revista *Bahia Análise & Dados*, estamos certos de que o seu conteúdo se reveste como mais um vetor que possibilitará a disseminação de mais conhecimentos em favor do aprimoramento da gestão pública, da revisão de metodologias de trabalho, da organização de equipes, da formação profissional e, consequentemente, trará maiores benefícios a toda sociedade.

REFERÊNCIAS

AMOROSO, Anastasia. *IA e Big Data*: novas ferramentas na luta contra a COVID-19. Disponível em: https://privatebank.jpmorgan.com/gl/pt/insights/investing/ai-and-big-data-new-tools-in-the-fight-against-covid-19. Acesso em: 8 dez. 2020.

BORTOLASSI, Leandro. Big Data auxilia o desenvolvimento de políticas públicas. *Exame*, São Paulo, 10 abr. 2019. Disponível em: https://exame.abril.com.br/blog/instituto-millenium/big-data-auxilia-o-desenvolvimento-de-politicas-publicas/. Acesso em: 10 set. 2020.

CASTRO, Sebastião Helvecio Ramos de. Utilização do Big Data no impacto de políticas públicas. *In*: CONFERENCIA IBEROAMERICANA DE COMPLEJIDAD, INFORMÁTICA Y CIBERNÉTICA, 8., 2018, Orlando. *Anais* [...]. Orlando: CICIC, 2018. Disponível em: http://www.iiis.org/CDs2018/CD2018Spring/papers/CB182EJ.pdf. Acesso em: 12 set. 2019.

EIGTH DATA SCIENCE INDEX. *Panorama da violência contra a mulher*. São Paulo: EIGTH, 2019. Disponível em: https://www.institutomillenium.org.br/images/news/PVCM-2019.pdf. Acesso em: 12 set. 2019.

FERRARI, Elisabete; AMORIM, Sonia Naves; TOMACHESKI, João Alberto. Sustentabilidade de iniciativas premiadas no concurso inovação: indícios de mudança da gestão no governo federal? *In*: CONGRESSOCONSAD DE GESTÃO PÚBLICA, 4., 2011, Brasília. *Anais* [...]. Brasília: CONSAD, 2011. Disponível em: https://repositorio.enap.gov.br/bitstream/1/736/1/Sustentabilidade%20de%20iniciativas%20premiadas%20no%20concurso%20inova%C3%A7%C3%A3o%20-%20ind%C3%ADcios%20de%20mudan%C3%A7a%20da%20gest%C3%A3o%20no%20governo%20federal.pdf. Acesso em: 13 set. 2019.

LIMA, Cláudio. Estratégias para a inserção de Machine Learning em políticas governamentais de inovação. *In*: SEMANA DE INOVAÇÃO EM GESTÃO PÚBLICA: TRANSFORMAÇÃO DIGITAL, 3., 2017, Brasília. *Anais* [...]. Brasília: Ministério do Planejamento, Desenvolvimento e Gestão: Escola Nacional de Administração Pública, 2017. Disponível em: http://inova.gov.br/wp-content/uploads/2018/03/Relat%C3%B3rio-3%C2%AA-Semana-de-Inova%C3%A7%C3%A3o-Plataforma.pdf. Acesso em: 13 set. 2019.

MENDONÇA, Fabrício Martins; DANTAS, Mário Antônio Ribeiro. Covid-19: where is the digital transformation, Big Data, artificial intelligence and data analytics?. *Revista do Serviço Público*, Brasília, v. 71, p. 212-234, 2020. Special edition. Disponível em: https://doi.org/10.21874/rsp.v71i0.4770. Acesso em: 5 dez. 2020.

SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA: INTELIGÊNCIA ARTIFICIAL: A NOVA FRONTEIRA DA CIÊNCIA BRASILEIRA, 17., 2020, Brasília. *Anais* [...]. Brasília: Ministérioda Ciência, Tecnologia, Inovações e Comunicações, 2020. Disponível em: https://doi.org/10.21874/rsp.v71i0.4770.

VOLPATO, Tiago; RUFINO, Ricardo Ribeiro; DIAS, Jaime William. *Big Data: transformando dados em decisões*. Paranavaí: Universidade Paranaense, 2014. Disponível em: https://docplayer.com.br/3437738-Big-data-transformando-dados-em-decisoes.html. Acesso em: 9 set. 2019.

WILSON DA SILVA GOMES

Comunicação, sociedade e big data

Wilson da Silva Gomes é um daqueles intelectuais, incomuns no país, que possuem o mérito de sair dos muros da universidade. Com um excelente currículo acadêmico, professor titular da Universidade Federal da Bahia (UFBA), pesquisador 1A do Cnpq, mestre e doutor em Filosofia pela Pontificia Università San Tommaso d'Aquino (Angelicum), além de graduado em Teologia (Universitas Gregoriana) – cogitou ser padre –, ambas em Roma, o professor publicou 12 livros, sendo o Transformações da Política na Era da Comunicação de Massa (Paulus. 2004), o mais vendido deles, e tornou-se referência também em espaços como as redes sociais. Em seus perfis no Facebook (@wilson.gomes.9883) e do Twitter (@willgomes), com milhares de seguidores cada, faz comentários diários sobre temas como democracia, redes, política e sociedade, mostrando um humor ácido e rara inteligência. Tendo os efeitos sociais dos usos da tecnologia como recente tema de sua pesquisa, foi convidado a participar de audiência da CPMI das fake news. Nesta entrevista, gentilmente concedida a nós, Gomes fala um pouco sobre o uso do big data e seus impactos na comunicação e sociedade atuais..

BA&D – O uso de Inteligência Artificial para trabalhar com grandes volumes de dados tem sido uma prática muito recente por empresas e governos para interferir no comportamento das pessoas, influenciando as suas decisões em situações como eleições e consumo de mercadorias. Como esses avanços tecnológicos podem ser utilizados para subsidiar as políticas públicas no Brasil?

Wilson Gomes – Acho que não vou chocar qualquer pessoa da área se disser que, na verdade, os maiores investimentos em IA foram feitos até hoje para "otimizar" a publicidade. Quer dizer, para aumentar o alcance e a eficiência da comunicação persuasiva de tipo comercial. A razão disso é basicamente porque os interesses comerciais é que põem o dinheiro na mesa para pagar mil desenvolvedores construindo os algoritmos de que necessitam no seu modelo de negócios. Por outro lado, é bom ter as coisas na justa proporção, pois tentar "influenciar decisões", políticas, eleitorais ou de consumo é coisa que se faz desde sempre e é o que também se chama de "persuasão" e de propaganda. A retórica antiga foi desenvolvida na Atenas democrática na convicção de que as pessoas podem convencer as outras e não há nada de errado nisso. Aliás, a democracia antiga surge quando se começou a achar que é melhor convencer do que forçar, melhor persuadir do que impor.

Por outro lado, influenciar decisões eleitorais não é algo que governos democráticos fazem ou devam fazer. Quem está autorizado socialmente a tentar tal coisa são partidos, organizações e movimentos políticos. E temos exemplos acumulados do uso de coleta e tratamento de rastros digitais em campanhas eleitorais, por meio de Inteligência Artificial, como o fizeram Barack Obama, os defensores do Brexit e Donald Trump, para ficarmos nos exemplos mais vistosos.

Imagino que a questão fundamental aqui seja como passar de um uso de IA em benefício de interesses privados ou particulares para um emprego em benefício geral da sociedade. E aí entrariam na equação, certamente, governos, organizações do terceiro setor e outros âmbitos do Estado que operam no interesse geral da democracia (Ministério Público, Judiciário, Legislativo). Claro, há muitas razões por trás, por exemplo, de uma política pública ou de um projeto de lei, muitas delas indiferentes a dados e evidências, pois orientadas pelo perde-e-ganha da luta partidária, das facções de poder ou simplesmente pelo interesse de grandes atores sociais por trás dos panos e por suas agendas ocultas. Por outro lado, sempre houve, igualmente, uma genuína preocupação com a coleta de informações e com a busca de evidências que pode se beneficiar com a analítica de grandes volumes de dados extraídos de rastros digitais. E é neste aspecto que se pode começar a pensar na IA a serviço de governos mais democráticos e mais progressistas.

Sempre houve, igualmente, uma genuína preocupação com a coleta de informações e com a busca de evidências que pode se beneficiar com a analítica de grandes volumes de dados extraídos de rastros digitais

Qualquer Estado e em qualquer época sempre precisou produzir, coletar, tratar e analisar dados. Sempre teve em sua posse, até os grandes processos de digitalização e datificação digital da vida dos últimos 20 anos, os maiores volumes de dados disponíveis em qualquer sociedade. Seja como for, um setor eficiente do Estado precisa do máximo de informações disponíveis, tratadas e analisadas da melhor maneira possível, para cumprir a sua função, seja ela a entrega permanente e universal de serviços públicos, a produção de políticas públicas baseadas em evidências, o processo legislativo, as incontáveis decisões executivas e judiciais. A diferença é que, neste momento da história, os nossos dados ou indicadores captados e tratados digitalmente permitem uma compreensão muito melhor, mais precisa e refinada das necessidades da população. Ou possibilitam fazer predições (inclusive para a tomada de providências no que incumbe ao Estado) muito mais precisas e segmentadas. Os dados digitais são mais volumosos a cada dia e, tratados com recursos de IA, são extremamente mais precisos e refinados (*fine-grained*).

Pouquíssimos são os cidadãos que não produzem voluntariamente rastros digitais em mídias e aparelhos digitais ou realizando as operações comuns do dia a dia

Em tese, é possível hoje ter perfis razoavelmente completos dos 209,5 milhões de brasileiros. Pouquíssimos são os cidadãos que não produzem voluntariamente rastros digitais em mídias e aparelhos digitais ou realizando as operações comuns do dia a dia. Menos ainda os que não são objeto de registro digital inadvertidamente. A IA está aí para ajudar a preencher as lacunas e para fazer as combinações de dados das mais variadas fontes (a fusão de dados). Processos de machine learning estão aí para ajudar a aprender os nossos padrões de comportamento, preferências e personalidade, bem como para permitir que se façam melhores predições e antecipações ou até para que se entenda as necessidades e padecimentos da população. Muitos dos dados são coletados e processados em tempo real, o que daria a quem os analisa uma capacidade impressionante de entender o que está acontecendo no País aqui e agora.

Tudo isso, naturalmente, pode ser usado para classificar a nossa personalidade de forma que uma campanha possa distribuir, por *micro-targeting*, mensagens feitas sob medida para nos convencer a comprar sapatos ou perfumes, a votar em x ou y, mas também poderiam servir para persuadir grupos específicos de pessoas resistentes às campanhas de saúde pública, como os mais jovens ou os mais bolsonaristas, por exemplo, durante a pandemia do coronavírus. A identificar, em tempo real e pela conversa digital, áreas da cidade atingidas por surtos de dengue ou a geolocalizar com precisão áreas e horários em que as pessoas estão mais sujeitas a assaltos ou níveis de poluição por bairro e rua. A prever desastres ambientais, ciclos sazonais de secas e cheias, velocidade da expansão de pandemias, necessidade de indução do Estado na formação de especialistas em tal ou qual área.

Em ambientes sociais, digitais ou não, você é incentivado à conformação e punido pela dissonância, você experimenta o amor do grupo ou o isolamento

A IA pode servir para a manipulação e o engano, mas também pode servir para dar *smartness*, inteligência assistida por computadores, à vida urbana e à demanda por serviços públicos. E muito disso já está operando. Devem estar chegando a 100% as decisões tomadas por IA, por exemplo, nas redes de distribuição elétrica. Há projetos-pilotos e rotinas envolvendo IA em várias iniciativas das chamadas cidades inteligentes e governos inteligentes (*smart-government*) ao redor do mundo, todos focados no uso eficaz dos recursos públicos e do capital humano, na condução de soluções eficientes e sustentáveis para problemas da coletividade, em melhorar a vida das pessoas.

De Big Data e IA, pode-se dizer o mesmo que da maioria dos recursos que as revoluções digitais e dos dados trouxeram consigo: o que se pode empregar para manipular e dominar também pode ser empregado para promover mais e melhores democracias. Tudo depende dos usos sociais que lhes são dados.

BA&D – Em que medida você acredita que a vida das pessoas já esteja sendo comandada pela via das redes sociais em mensagens cifradas e direcionadas a grupos focais utilizando Inteligência Artificial e Big Data?

WG – "Comandada" me parece demais. "Influenciada" talvez seja um termo mais adequado. A resposta é que em boa medida tal coisa pode estar acontecendo, mas isso, por outro lado, não quer dizer nada de extraordinário.

As mídias digitais, e não apenas as redes sociais digitais, configuram hoje os mais relevantes ambientes sociais e a forma principal da esfera pública política ou cultural. Ambientes sociais são contextos de convivência em que circulam e se consomem informações (no caso, "conteúdos"), em que as pessoas se expressam, consideram o que os outros dizem e são consideradas no que manifestam, em que expressões, atitudes e comportamentos estão expostos às punições e recompensas dos grupos de referência. Em ambientes sociais, digitais ou não, você é incentivado à conformação e punido pela dissonância, você experimenta o amor do grupo ou o isolamento. O medo do isolamento, alguns autores dizem, é o recurso por meio do qual o coletivo passa a ter poder sobre mim. Influência, portanto, é parte fundamental da nossa experiência social em ambientes de referência, seja a família, os amigos, os companheiros de escola, os colegas de trabalho ou a Igreja. Inclusive nos ambientes sociais digitais que são formados por outras afinidades, como preferências políticas ou interesses em comum.

Pode-se imaginar, entretanto, que além dos processos espontâneos (ou "orgânicos", como se diz nas plataformas digitais) de influência, possam estar em curso processos de influência dirigidos intencio-

nalmente. Isto porque os nossos rastros digitais, capturados e tratados, permitem que nós e as nossas redes de contatos, amigos e influências sejamos "perfilados", ou seja, que se descubra os nossos padrões de preferências, comportamentos e inclinações. E, com isso, nós e nossos ambientes sociais nos tornamos previsíveis. Como já se demonstrou, com 100 likes em posts publicados em uma plataforma digital, métodos de *machine learning* conseguem prever a personalidade de uma pessoa de modo consideravelmente preciso. Ora, se eu sei que o ambiente social digital x adota o padrão y de convicções, crenças, inclinações e preferência, é possível influenciá-lo por meio de mensagens sob medida; se 50 mil perfis são profundamente impactados por três influenciadores digitais cujo perfil psicológico foi aprendido pela máquina, é possível exercer efeito indireto sobre 50, 500 mil pessoas, produzindo mensagens específicas e bem refinadas para um número muito pequeno desses influenciadores.

Com 100 likes em posts publicados em uma plataforma digital, métodos de *machine learning* conseguem prever a personalidade de uma pessoa de modo consideravelmente preciso

Além disso, é possível não apenas encontrar os padrões de medos, de esperanças ou de raiva das peesoas neste momento em determinado contexto social por meio de IA aplicada à Big Social Data (grandes volumes de dados provenientes de mídias digitais), como também classificar 200 milhões de pessoas em *clusters* a partir dos sentimentos a que são mais suscetíveis. A partir daí, é também possível prover gatilhos para medos, esperanças e ódio em conformidade com intenções políticas determinadas, por exemplo, e dar um jeito de que essas mensagens-gatilho sejam empacotadas sob medida e entregues de forma customizada, de um jeito ou de outro, a cada uma das pessoas que se queira alcançar.

Por fim, e só para ficarmos em pouquíssimos exemplos, é possível descobrir padrões políticos (ou de consumo) muito mais refinados do que os que são capazes de fazer a nossa humana capacidade analítica. E de se criar algoritmos para lidar com eles. Somos, por exemplo, habituados a distribuir as inclinações políticas das pessoas por duas dimensões ou espectros. Primeiro, o espectro ideológico entre direita, centro e esquerda. Depois, o espectro moral entre conservadores, moderados e liberais (ou progressistas). Podemos acrescentar eventualmente outras dimensões como a contraposição entre pragmáticos e idealistas, entre estatistas e antiestatistas, entre personalidades flexíveis e personalidades autoritárias, entre secularistas e religiosos etc. De toda sorte, há um limite de aplicação humana dessas dimensões a um conjunto limitado de perfis, depois de uma coleta penosa e custosa de informações sobre cada pessoa, por meio de *surveys*. Ora, com IA se pode, em princípio, usar não duas ou dez dimensões para classificar as personalidades políticas, mas, digamos, 500, de maneira cada vez mais precisa. E o *machine learning* nos permitirá fazer predições muito exatas sobre atitudes e comportamentos

Primeiro, claro, a revolução que consistiu na digitalização da vida, em praticamente todas as suas dimensões

das pessoas submetidas a certas mensagens sob medida, provenientes de determinadas fontes confiáveis e com dada intensidade.

Isso não quer dizer, contudo, que tudo isso seja feito ou seja feito o tempo todo. Não são processos baratos e há sempre a questão de quem vai colocar o dinheiro na mesa para pagar por tudo. Mas quer dizer que é possível acontecer e que, portanto, se é possível, é também plausível imaginar que, pelo menos eventualmente, realmente aconteça.

BA&D – Hoje, vemos o "boom" das redes sociais e sua capacidade de ampliação dos discursos, a promover e democratizar debates, mas também criando espaços de "bolha". Em que medida as análises de Big Data de redes sociais – a exemplo da feita pela DAPP, da Fundação Getúlio Vargas, que mede as interações no Twitter – valem como termômetro da opinião pública?

WG – Acho que o termo "boom" não se aplica mais ao caso das redes sociais digitais, em particular, nem à mídias sociais, em geral. O "boom" se refere a uma irrupção, um crescimento intenso, rápido e repentino. Em 2005 e 2006, houve algo que se poderia descrever assim, mas um fenômeno que continua crescendo consistentemente por mais 15 anos e que já foi completamente incorporado ao panorama da digitalização das nossas vidas não pode mais ser descrito nesses termos. Isto posto, é claro que a política em ambientes e baseada em recursos digitais é um fato característico da nossa experiência política hoje, com impactos enormes nas eleições e na formação e manutenção da opinião pública.

O fato é que passamos por duas revoluções recentes, com consequências que ainda estamos longe de avaliar e evoluções que não conseguimos prever e que afetaram enormemente as socidades contemporâneas. Primeiro, claro, a revolução que consistiu na digitalização da vida, em praticamente todas as suas dimensões: nas interações sociais, na produção e consumo do conhecimento, nas relações íntimas e privadas, na interação com o Estado, no trabalho e na produção econômica, no comércio, na política. Simplesmente não há âmbito da vida que não tenha sofrido níveis importantes de mediação digital.

Segundo, sobre a digitalização se realiza uma segunda revolução que é a datificação digital da vida. Isto significa que, por um lado, nos tornamos produtores involuntários de volumes gigantescos de dados. Os nossos rastros digitais se multiplicam e são coletados, tratados e usados. Claro, vale notar a este ponto que os dados digitais não são importantes por serem volumosos, mas por serem significativos, mas volume, diversidade de fontes e a sua origem inadvertida são relevantes, sim.

Ora, há muitas instituições buscando ativamente combinar dados de múltiplas fontes para deles extrair significados e para, a partir deles, produzir aplicações de interesse particular ou público. Os dados e a inteligência de dados (humana, assistida por máquinas ou maquínica) se tornaram os ativos mais importantes para a terceira década do século XXI, e é aí que se encaixa um tipo específico de dados, provenientes dos nossos rastros digitais na "internet social", na internet de interações humanas, redes e outras formas de capital social.

Neste âmbito, há duas aplicações de IA a "Big Social Data" que chamam a atenção. A primeira é resultante da ambição de previsão de comportamentos sociais e políticos. Umas das áreas com mais registros de tentativas de desenvolvimento de algoritmos de predição é a da previsão de resultados eleitorais a partir dos climas de opinião ou aumento de interações em plataformas de redes sociais ou pelos fluxos de buscas no Google, por exemplo. A segunda é a analítica de grandes volumes de dados para a detecção do estado da opinião e do sentimento públicos, tentando entender para onde se está movendo a percepção pública dos fatos políticos ou sociais, quais são as suas inflexões e como público reage a temas, fatos, projetos e declarações políticas.

Quem tem que tornar as pessoas menos vulneráveis, seja a controles sociais, seja à exploração de dados do indivíduos para fins comerciais e objetivos escusos, é a lei, não as habilidades individuais

BA&D – Temos altas taxas de analfabetismo e uma relativa baixa escolaridade no Brasil, notadamente no Nordeste. Você considera que o fator educacional constitui uma vulnerabilidade para um forte controle social, agora intensificado pelos avanços tecnológicos de Big Data?

WG – Quem tem que tornar as pessoas menos vulneráveis, seja a controles sociais, seja à exploração de dados do indivíduos para fins comerciais e objetivos escusos, é a lei, não as habilidades individuais. Para isso existem os três poderes e os seus controles recíprocos, existe a norma constitucional e os instrumentos legais e de aplicação da Lei, de que o Estado liberal-democrático deveria ser dotado. Por isso, a sociedade correu atrás de uma Lei Geral de Proteção de Dados e há um esforço mundo afora para construir instituições independentes e normas que protejam as formas de exploração de dados pessoais que impliquem controle ou manipulação por parte dos atores socialmente poderosos ou do próprio Estado.

Então, não, não há nem pode haver correlação entre nível de escolarização e controle social. Mesmo porque sociedades de controle se estabelecem onde não há constituições liberais em que as liberdades individuais sejam garantidas por um Judiciário independente; e isso não tem correlação com PIB, IDH ou escolaridade. Não nos faltam exemplos de sociedades ricas e altamente escolarizadas que são claramente sociedades de controle dos seus cidadãos.

BA&D – A Bahia tem características muito peculiares no comportamento, ainda se utilizando fortemente das relações pessoais presenciais. Diante disso, estratégias de comunicação de massa podem conduzir ou influenciar hábitos de compra e consumo menos saudáveis?

WG – Não consigo imaginar que a Bahia tenha peculiaridades comportamentais distintas de outras regiões do País e do mundo. Mesmo porque pode haver bastante pessoalidade e intensa presença em relações pessoais mesmo quando estas são digitalmente mediadas. A ideia de que vivemos em dois universos que correm em paralelo, um em que se vive por imersão corpórea e outro por meio de máquinas conectadas, de forma que as pessoas mais atualizadas tecnologicamente se moveriam constantemente entre um universo e outro, foi típica do primeiro momento de popularização da internet no Brasil, entre 1995 e 1999. Nesta época, era comum se fazer contraposições binárias entre o real e o "virtual" e se caracterizar a experiência da conexão digital como entradas e saídas "da internet". Ao se entrar na internet, saía-se do mundo, claro. Do mesmo modo que era comum fazer um rol de perdas à medida que certos tipos de pessoas pareciam preferir amigos, conversas e sexo "virtuais" em vez das alternativas "reais".

Não é preciso entrar nem sair de coisa alguma, uma vez que o estado de conexão permanente [...] onde estivermos, haverá conexão

Por volta de 2010, essa mentalidade e esse vocabulário deixaram inteiramente de fazer sentido. Primeiro, desfez-se a ideia binária das dupla existência: não há dois mundos, mas um só, com múltiplos recursos, aos quais acrescentamos também os recursos digitais. Não é preciso entrar nem sair de coisa alguma, uma vez que o estado de conexão permanente, a fase dos aparelhos que não precisam mais ser desligados, já havia se popularizado na década anterior. Além disso, a terceira fase da vida digital, aquela baseada em dispositivos móveis e aplicativos, tranformava a experiência digital em algo ubíquo: onde estivermos, haverá conexão. Estamos aqui agora no bar com certos amigos e ao mesmo tempo, sem qualquer sentido obrigatório de descontinuidade, com outros amigos, familiares e até com chefes e colegas de trabalho por meio das janelas digitais que me dão dispositivos móveis conectados. Estamos aqui no trabalho e ao mesmo tempo consumindo conteúdo digital, interagindo em ambientes digitais e colaborando com fluxos incessantes de dowloads e uploads. Não existem mais vida off-line e vida on-line, é a mesma vida, on e off, simultaneamente.

Assim, a digitalização e a datificação da vida não competem com relações presenciais. Elas transformam as relações presenciais e as levam a uma outra dimensão: da ubiquidade, da conexão permanente, da superação das amarras do tempo e do limite do espaço. O que seria da nossa sanidade mental e da nossa qualidade de vida se nesta pandemia, em que boa parte da população foi forçada abruptamente e por uma duração tão grande, a separar-se de famílias, colegas,

ambientes de trabalho, amigos, amores e amantes, se a vida de todo mundo já não fosse tão intensamente digitalizada?

Este é o quadro do mundo, não apenas da Bahia, e não temos razão para crer que no nosso estado as coisas tenham ocorrido diferentemente. Então, sim, é claro que vivemos sob as mesmas circunstâncias de qualquer população que passou pela revolução digital e pela subsidiária revolução dos dados digitais. Nessas circunstâncias há espaço para pensar que se possa conduzir ou influenciar hábitos de compra e consumo menos saudáveis. Não há razão, contudo, para se pensar que a indução de hábitos ou decisões de consumo de produtos seja a única coisa nem a principal coisa que se possa fazer, nem que os produtos consumidos sejam os menos saudáveis. Nas novas condições da vida com alto poder de fogo digital, as coisas podem ser mais complexas.

A ideia de uma manipulação unidirecional baseada em Big Data, servindo-se de Inteligência Artificial e voltada para alcançar o seu *target* em ambiente digital, descreve de forma imperfeita um dos lados da questão

Admitidas a digitalização e a datificação da vida social agora, no limiar da terceira década do século XXI, deve-se considerar que pessoas podem ter atitudes e comportamentos influenciados, em ambientes digitais ou por meio de aplicativos, por recursos baseados em analítica de Big Data e Inteligência Artifical, sim, mas quem disse que elas também não podem produzir impactos sobre marcas, empresas e produtos usando a internet social para atividades de detratação, boicote e cancelamentos?

Organizações e marcas podem usar analítica de Big Data para descobrir o que cada pessoa gosta ou precisa, podem usar Inteligência Artificial para induzir climas de opinião e manipular opiniões, atitudes e desejos, mas as pessoas e grupos podem embaralhar o jogo, por meio de tags e campanhas, para que as empresas saibam que os seus sentimentos mudaram com relação a uma determinada marca ou partido político e que ou elas mudam ou enfrentarão as consequências de andarem em desalinho com certa nova sensibilidade social.

Isso me lembra de uma crônica de Umberto Eco, em que se conta de um cachorro russo que andava se gabando com os seus colegas caninos, explicando-lhes como havia conseguido condicionar um cientista chamado Pavlov para que ele o alimentasse com guloseimas. Isso para dizer que a ideia de uma manipulação unidirecional baseada em Big Data, servindo-se de Inteligência Artificial e voltada para alcançar o seu *target* (atitudes e comportamento das pessoas) em ambiente digital, descreve de forma imperfeita um dos lados da questão. Que pessoas e grupos são também muito conscientes da partida que se está jogando em ambientes digitais e de como é possível usar artifícios digitais para impactar organizações, instituições, marcas e atores sociais, atacando imagens e credibilidade, fixando na imperecível memória social digital os registros das condutas condenadas, conduzindo fluxos de opinião pública adversária e, portanto, desmantelando anos de trabalho de construção de imagem e reputação.

Nesta edição da *Bahia Análise & Dados* trazemos um Especial com um estudo de caso sobre a evolução da taxa de letalidade por covid-19 no Maciço de Batrité, no Ceará.

Trata-se do resumo de uma dissertação, defendida e gentilmente cedida pelo pesquisador Erivando de Sena Ramos.

O presente trabalho científico, cujo conteúdo apresenta correlação com o tema da revista, objetiva conhecer, identificar e analisar padrões comportamentais da evolução de taxa de letalidade por coronavírus (covid-19) e casos confirmados entre municípios da localidade citada.

Trata-se de um estudo populacional, ecológico-descritivo, transversal e de série temporal. No período selecionado para a investigação, constatou-se que Baturité é o município mais densamente povoado, com menos casos confirmados; entretanto, com uma maior taxa de letalidade por coronavírus (5,24%) constatada no dia 17 de setembro de 2020. Por outro lado, o Acarape é menos densamente povoado, com mais casos confirmados; contudo, tem a menor taxa de letalidade por coronavírus (0,83%). Importa ressaltar que é o único município sem sinal de estabilização de casos confirmados até a data referenciada.

O conhecimento, a identificação e a análise de padrões da evolução de taxa de letalidade por coronavírus e casos confirmados são decisivos para determinar o ajustamento das estratégias de mitigação e permitir o planejamento das necessidades de cuidados de saúde conforme a desenvoltura da epidemia.

A evolução da taxa de letalidade e casos confirmados de covid-19 entre municípios do Maciço de Baturité, no Ceará

ERIVANDO DE SENA RAMOS

Mestre em Sociobiodiversidade e Tecnologias Sustentáveis, pela Universidade da Integração Internacional da Lusofonia Afro-Brasileira (Unilab) e especialista em Ciência de Dados e Big Data Analytics, pela Universidade Estácio de Sá (Unesa). erivandoramos@unilab.edu.br

JOHN HEBERT DA SILVA FELIX

Doutor e mestre em Engenharia de Teleinformática pela Universidade Federal do Ceará (UFC). Professor associado do Instituto de Engenharias e Desenvolvimento Sustentável, professor permanente do Mestrado Acadêmico em Sociobiodiversidade e Tecnologias Sustentáveis (MASTS) e do Programa de Pós-Graduação em Energia e Ambiente (PGEA).

FRANCISCO HORÁCIO DA SILVA FROTA

Pós-doutor, pela Universidade de Coimbra e doutor em Sociologia Política, pela Universidad de Salamanca (USAL). Professor do Programa de Pós-graduação em Sociologia, professor e coordenador do Programa de Pós-graduação em Políticas Públicas, ambos da Universidade Estadual do Ceará (UECE).

MARIA HELENA DE PAULA FROTA

Doutora em Sociología, pela Universidad de Salamanca (USAL) e mestre em Sociologia, pela Universidade Federal do Ceará (UFC). Professora adjunta da Universidade Estadual do Ceará (UECE), pesquisadora do Núcleo de Pesquisas Sociais (Nupes) da UECE, representante institucional Secretaria da Ciência, Tecnologia e Educação Superior (Secitece)/UECE no Conselho Cearense dos Direitos da Mulher (CCDM).

ESTE TRABALHO científico, cujo conteúdo apresenta correlação com o tema desta edição da revista *Bahia Análise & Dados*, objetiva conhecer, identificar e analisar padrões comportamentais da evolução de taxa de letalidade por coronavírus (covid-19) e casos confirmados entre municípios do Maciço de Baturité-Ceará – no período de 15/04/2020 a 17/09/2020. Trata-se de um estudo populacional, ecológico-descritivo, transversal e de série temporal.

No período selecionado para a investigação, constatou-se que Baturité é o município mais densamente povoado, com menos casos confirmados; entretanto, com uma maior taxa de letalidade por coronavírus (5,24%) no dia 17/09/2020. Por outro lado, o Acarape é menos densamente povoado, com mais casos confirmados; contudo, tem a menor taxa de letalidade por coronavírus (0,83%). Importa ressaltar que é o único município sem sinal de estabilização de casos confirmados até o dia 17/09/2020.

O conhecimento, a identificação e a análise de padrões da evolução de taxa de

Está acontecendo a mais grave crise sanitária no mundo neste primeiro quarto do século XXI, com a pandemia da covid-19 associada ao novo coronavírus (SARSCoV2)

letalidade por coronavírus e casos confirmados são decisivos para determinar o ajustamento das estratégias de mitigação e permitir o planejamento das necessidades de cuidados de saúde conforme a desenvoltura da epidemia.

O tema reveste-se de grande relevância neste momento delicado vivenciado pela sociedade em âmbito global e esperamos contribuir com este estudo, para o maior aprofundamento das análises e a robustez desta publicação inovadora da Superintendência de Estudos Econômicos e Sociais da Bahia sobre Big Data e Políticas Públicas.

Está acontecendo a mais grave crise sanitária no mundo neste primeiro quarto do século XXI, com a pandemia da covid-19 associada ao novo coronavírus (SARSCoV2). Observam-se grandes conflitos nas rotinas e dinâmicas dos indivíduos e de toda a sociedade, bem como sobrecarga do sistema de saúde, mortes, profunda crise econômica, instabilidade política e conflitos institucionais.

Os sinais e sintomas característicos da doença são febre, tosse, dificuldade para respirar (BRASIL, 2020a) além de fadiga, mialgia, congestão nasal, coriza, espirros, dor de garganta, dor de cabeça, tontura, náusea, vômito, dor abdominal, diarreia (RECOMMENDATIONS..., 2020) e pneumonia atípica (NUNES *et al.*, 2020). O novo coronavírus pertence à família *coronaviridae*, que tem outros membros com

semelhanças filogenéticas, incluindo SARS-CoV, que causa Síndrome Respiratória Aguda Grave (SARS), e MERS-CoV, que causa Síndrome Respiratória do Oriente Médio (MERS), (BRASIL, 2020a; WU *et al.*, 2020). A maioria das infecções provocadas pelo novo coronavírus é de baixa patogenicidade, entretanto, pode eventualmente levar a infecções graves em pacientes imunodeprimidos, bem como afetar especialmente crianças, pessoas com comorbidades e idosos (BRASIL, 2020b).

Apesar disso, existem mais quatro tipos de coronavírus conhecidos por originar doenças nos indivíduos (229E, OC43, NL63, HKU1), mas a patologia é menos grave, causando resfriado comum como sintoma (LIMA, 2020).

O novo coronavírus foi detectado pela primeira vez em 31 de dezembro de 2019, em Wuhan, na China (LANA *et al.*, 2020), mas sua circulação foi confirmada apenas em 9 de janeiro de 2020. Em 16 de janeiro, foi notificado um caso importado da doença em território japonês e, em 21 de janeiro, foi reportado o primeiro caso importado nos Estados Unidos. Em 30 de janeiro de 2020, a OMS declarou que a epidemia era uma emergência internacional (WORLD HEALTH ORGANIZATION, 2020c). No final de janeiro, vários países já haviam confirmado importações de

caso, incluindo Canadá e Austrália. Mas outros países, como Coreia do Sul, Itália, Irã, Japão, França e Alemanha, já possuíam um número significativo de casos confirmados (PIRES *et al*., 2020).

O primeiro caso confirmado de coronavírus no Brasil ocorreu no estado de São Paulo, no dia 26 de fevereiro de 2020

No Brasil, em 7 de fevereiro, havia nove (9) casos em investigação, mas sem registros de casos confirmados (WORLD HEALTH ORGANIZATION, 2020d). O primeiro caso confirmado de coronavírus no Brasil ocorreu no estado de São Paulo, no dia 26 de fevereiro de 2020 (RODRIGUES, 2020). No estado do Ceará, de acordo com informações da Secretaria da Saúde do Estado do Ceará (Sesa) (SESA, 2020), a confirmação dos três primeiros casos positivos veio no dia 15 de março de 2020. E, no dia 20 de março de 2020 (CEARÁ, 2020b). por força de um decreto legislativo, foi estabelecido o início do isolamento social no estado do Ceará (CEARÁ, 2020a).

Em até 18 de junho de 2020, foram registrados casos em mais de 181 países, com 8.242.998 infectados confirmados e 445.535 mortos (NUNES *et al*., 2020) e com uma taxa de letalidade de 5,4%. No Brasil, até a mesma data, a taxa de letalidade era de 4,9% (47.748 mortes em 978.142 casos oficiais de infecção), mais baixa que a média dos 181 países supracitados, contudo, alta para o padrão esperado em estudo anterior da Organização Mundial da Saúde (OMS) (WORLD HEALTH ORGANIZATION, 2020b), de cerca de 2% (LOVELACE JR.; HIGGINS-DUNN, 2020).

A covid-19 não se comporta como uma gripe qualquer. Sabe-se bem sobre a gripe sazonal, por exemplo, sobre como ela é transmitida e quais tratamentos funcionam para suprimir a doença, mas essa mesma informação ainda é questionada quando se trata do coronavírus (LOVELACE JR.; HIGGINS-DUNN, 2020). O novo coronavírus não está transmitindo exatamente da mesma forma que a gripe, e é uma doença para a qual não temos vacina[1] nem tratamento e cuja transmissão ainda não compreendemos por completo. Não entendemos totalmente a mortalidade de casos, mas o que nos entusiasma genuinamente é que, ao contrário da gripe, contra a qual os diferentes países lutaram e implementaram medidas fortes, notamos que o vírus é suprimido, de acordo com Ryan[2] (LOVELACE JR.; HIGGINS-DUNN, 2020).

Em 11 de março, a Organização Mundial da Saúde declarou que o novo coronavírus é uma pandemia global (PIRES *et al*., 2020).

Em face da emergência da pandemia originada pelo vírus SARS-CoV-2, este estudo se justifica por possibilitar informações que podem auxiliar

1 No período do estudo, entre 15/04/2020 a 17/09/2020, as vacinas ainda não estavam disponíveis à população.

2 Diretor-executivo do Programa de Emergências de Saúde, da Organização Mundial de Saúde (OMS).

A pandemia da covid-19 foi declarada pela Organização Mundial da Saúde em março e ultrapassou a marca de 1 milhão de infectados e 150 mil mortos no Brasil após sete meses do primeiro caso

no planejamento de estratégias para se enfrentar o atual cenário (MARINELLI *et al.*, 2020).

O ritmo de cada país é diferente. Japão, Hong Kong e Singapura viram crescer as infecções de maneira paulatina desde janeiro. Em outros países, como Espanha, França e Alemanha, os casos dispararam seguindo o rastro da Itália. Mas a razão disso não é só a existência de infecções; há também um aumento nas detecções porque os países reforçaram seus protocolos. "É provável que essa mudança tenha tido um grande efeito no número de casos. A transmissão da doença pode ser alta, mas não é plausível que seja tão alta a ponto de gerar os picos que vimos em muitos países", diz Adam Kucharski, professor da London School of Hygiene & Tropical Medicine (PIRES *et al.*, 2020).

Uma forma de tentar comparar o ritmo do vírus em cada país é ver sua evolução desde que os primeiros casos foram confirmados (PIRES *et al.*, 2020).

REFERENCIAL TEÓRICO

A pandemia da covid-19 foi declarada pela Organização Mundial da Saúde em março e ultrapassou a marca de 1 milhão de infectados e 150 mil mortos no Brasil após sete meses do primeiro caso.

O novo coronavírus está se espalhando velozmente em quase todos os países, tendo contaminado pelo menos 38.394.169 indivíduos e matado 1.089.047 até o dia 15 de outubro de 2020. Mais da metade dessas mortes ocorreu no continente americano, que engloba o Brasil (596.312 mortes). Na sequência, vem o continente europeu, com pouco mais de 1/4 dessas mortes (251.478 mortes). As áreas do Pacífico Ocidental e da África são menos afetadas, tendo, respectivamente, 14.527 e 27.904 mortes (ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE, 2020).

A pandemia do coronavírus atingiu mais de 39.223.100 pessoas, de acordo com as contagens oficiais. Até o dia 16/10/2020, pelo menos 1.102.000 pessoas morreram e o vírus foi detectado em quase todos os países (THE NEW YORK TIMES, 2020).

O vírus continua afetando todas as regiões do mundo, mas alguns países estão enfrentando altas taxas de infecção, enquanto outros parecem ter controlado o vírus. Em lugares como República Checa, Bélgica, Países Baixos, Estados Unidos, França, Reino Unido, Espanha, Portugal e outros foram registrados, no dia 16/10/2020, números mais altos de casos novos, com uma média diária de pelo menos quatro novos casos por 100 mil indivíduos na semana anterior (THE NEW YORK TIMES, 2020).

Há países onde o número de novos casos é maior, mas está diminuindo, como ocorre em Israel, Brasil, Peru, Chile e Índia (THE NEW YORK TIMES, 2020). Também existem países, tais como Azerbaijão, África de Sul, El Salvador, Sérvia e Noruega, onde o número de novos casos é menor, mas está aumentando (THE NEW YORK TIMES, 2020).

Até essa data, havia no Brasil 153.214 mortes, um aumento de cerca de 31% nos três primeiros meses após junho (RIBEIRO, 2020). A taxa de letalidade brasileira, no dia 16/10/2020, era de 2,9%.

Na China, até essa data (dia 16 de outubro de 2020), foram registrados 91.436 casos confirmados de covid-19, com 4.746 óbitos (WORLD HEALTH ORGANIZATION, 2020a), e uma taxa de letalidade de 5,2%, cerca de 2,3 pontos percentuais a mais que a taxa do Brasil. Contudo, na China houve cerca de 61.778 óbitos a menos que no Brasil.

Na China, até essa data (dia 16 de outubro de 2020), foram registrados 91.436 casos confirmados de covid-19, com 4.746 óbitos [...], e uma taxa de letalidade de 5,2%, cerca de 2,3 pontos percentuais a mais que a taxa do Brasil. Contudo, na China houve cerca de 61.778 óbitos a menos que no Brasil.

A Itália contabiliza, até essa data, cerca de 391.611 casos positivos confirmados e 36.427 óbitos, tendo uma taxa de letalidade de 9,3% e superando a China em mais de 4 pontos percentuais (ITALIA, 2020).

Até essa data, houve, no estado do Ceará, 9.199 mortes em 263.143 casos confirmados. Assim, a taxa de letalidade ficou em 3,5%, acima da média brasileira, contudo, com 1,7 e 5,8 pontos percentuais a menos do que a da China e a da Itália, respectivamente.

Os casos em todo o mundo se estabilizaram em abril, depois que medidas de distanciamento social foram implementadas em muitas das áreas com surtos iniciais.

Mas, à medida que os países começaram a reabrir as atividades, em maio e junho, os Estados Unidos não conseguiram conter o ressurgimento da doença, tornando-se um dos principais responsáveis pelo aumento do número de casos em todo o mundo. Muitos países da América do Sul também estão enfrentando altas taxas de infecção, e os países europeus que tiveram surtos graves e iniciais estão vivendo um segundo aumento nos casos (THE NEW YORK TIMES, 2020).

O número de casos conhecidos de coronavírus nos Estados Unidos continua crescendo. Até o dia 16/10/2020, pelo menos 8.087.100 pessoas – em todos os estados, incluindo Washington D.C. e quatro territórios dos EUA – testaram positivo para o vírus, de acordo com um banco de dados do New York Times, e pelo menos 218.400 pacientes com o vírus morreram (THE NEW YORK TIMES, 2020). Isso configura uma taxa de letalidade de 2,7%, pouco abaixo da do Brasil (2,9%), porém com número maior de mortes por coronavírus que o Brasil.

Parece claro que o colapso dos sistemas de saúde e milhões de mortes dizimariam os países financeiramente e também na condição de sociedade, por isso salvar vidas humanas deve ser a primeira prioridade dos governos

A epidemia de covid-19 fornece evidências de que a urbanização e a globalização transformaram a forma como as pessoas convivem nas comunidades, e os avanços nos transportes e nas comunicações induziram uma rápida disseminação de doenças, tanto via meios de transporte doméstico quanto por rotas internacionais, como ônibus, trens, barcos e aviões. O aumento da densidade de pessoas em residências, transporte público, ambientes de trabalho, *shoppings centers* e eventos culturais, políticos, esportivos e religiosos ampliou a probabilidade de difusão do vírus (MORAES *et al*., 2020; VILLELA, 2020).

Nesse sentido, as desigualdades sociais induzem riscos mais elevados de contaminação, sobretudo em países de renda média e baixa, que, comumente, têm sistemas de saúde mais frágeis e uma competência limitada para lidar com um rápido aumento de casos. A pobreza contribui para a disseminação de epidemias e pandemias e, como não há medicamentos ou vacinas específicas disponíveis no período analisado, a mitigação da disseminação exponencial de covid-19 depende de estratégias de mitigação da comunidade (VILLELA, 2020).

É cada vez mais consensual que o isolamento social para prevenir a propagação do vírus é a estratégia correta não só do ponto de vista dos direitos humanos, mas também do ponto de vista econômico. Parece claro que o colapso dos sistemas de saúde e milhões de mortes dizimariam os países financeiramente e também na condição de sociedade, por isso salvar vidas humanas deve ser a primeira prioridade dos governos.

Conquanto a pandemia seja um fenômeno global, seu impacto é amplamente moldado por decisões tomadas por governos individualmente. Vários governos responderam rapidamente e outros, após uma resposta inicial lenta, reconhecem seu erro e adotam as recomendações da Organização Mundial da Saúde (OMS). Ressalva-se, contudo, que alguns governos ainda continuam ignorando as recomendações da OMS sobre como evitar a contaminação por coronavírus (SILVA; ARBILLA, 2020).

No Brasil, de acordo com o Ministério da Saúde (MS), até o dia 31 de março, foram confirmados 5.717 casos de covid-19, com 201 óbitos (LIMA-COSTA; BARRETO, 2020). A região Nordeste concentrou 875 (15,3%) desses casos, sendo a segunda região do País em número de casos, superada apenas pela região Sudeste.

O Nordeste, região mais afetada pela epidemia do novo coronavírus, é composto por nove estados: Alagoas, Bahia, Ceará, Maranhão, Paraíba, Pernambuco, Piauí, Rio Grande do Norte e Sergipe (MORAES *et al*., 2020).

O isolamento social foi pregado por todos os estados nordestinos desde o início da pandemia (AQUINO *et al*., 2020), mas apenas em algumas cidades

foi decretado o *lockdown*, como ocorre em Fortaleza e São Luís (MADEIRO, 2020), cidades onde o curso da epidemia se tornou mais inquietante.

Alagoas decretou o estado de calamidade pública cerca de uma semana após o anúncio da pandemia (MORAES *et al*., 2020). Desde então, o governo local somou 700 leitos clínicos e 300 leitos de terapia intensiva até 15 de maio de 2020 e acelerou a entrega de hospitais em construção na capital e no interior (ALAGOAS, 2020). Foi perceptível a queda da letalidade nas últimas semanas do período analisado. Contudo, essa queda pode não estar relacionada a tal investimento em sua estrutura sanitária, mas sim à ampliação da testagem da população e à subnotificação de óbitos (PRADO *et al*., 2020). Nesse sentido, ressalva-se que, dentro do recorte temporal, Alagoas apresentou letalidade menor que o parâmetro de 3,4% da OMS (WORLD HEALTH ORGANIZATION, 2020b), todavia, o valor varia conforme o país. Estudos demonstram que, epidemiologicamente, homens com idade entre 41 e 58 anos representam a grande maioria dos casos de pacientes confirmados (UNA-SUS, 2020), sendo febre e tosse os sintomas mais presentes (MARINELLI *et al*., 2020).

A taxa de letalidade na China está em torno de 3,8% (BRASIL, 2020b). Ainda assim, a maioria dos estados nordestinos registrou letalidade acima da observada para o Brasil (5,4%), principalmente os estados de Piauí e Pernambuco. Isso sugere a falta de capacidade de atendimento adequado aos casos iniciais, seja por falha na suspeição, na notificação, no diagnóstico laboratorial ou no aparato do atendimento oportuno de cuidados intensivos. Além disso, há baixa cobertura de testagem (BRASIL, 2020b; MARINELLI *et al*., 2020).

Estudos apontam a evolução dos casos de covid-19 no Brasil e conhecimentos prévios de outras emergências de saúde que estabeleceram importante legado no enfrentamento das epidemias e evidenciaram a capacidade científica do País (CRODA *et al*., 2020).

A despeito dos impactos negativos, a pandemia do coronavírus criou novos ensejos, despontou exemplos de solidariedade nas comunidades locais e permitiu o compartilhamento de recursos, informações e experiências de países que estão em estágio mais avançado da pandemia ou com melhores resultados e conhecimentos sobre o controle da propagação. As comunidades científicas em todo o mundo deram as mãos e muitas universidades organizaram grupos de pesquisadores e estudantes para ajudar nos esforços de lenitivo da pandemia de todas as formas imagináveis (SILVA; ARBILLA, 2020).

A pandemia é uma forte advertência de que, para estarmos preparados para o futuro, é necessária uma mudança em nossa mentalidade, compromissos e valores (SILVA; ARBILLA, 2020).

A pandemia é uma forte advertência de que, para estarmos preparados para o futuro, é necessária uma mudança em nossa mentalidade, compromissos e valores

A economista britânica Kate Raworth, da Universidade de Oxford, corretamente mencionou que o desafio da humanidade no século XXI é atender às necessidades de todos dentro das possibilidades do planeta

A economista britânica Kate Raworth, da Universidade de Oxford, corretamente mencionou que o desafio da humanidade no século XXI é atender às necessidades de todos dentro das possibilidades do planeta (RAWORTH, 2017).

PROCEDIMENTOS METODOLÓGICOS

Este estudo populacional, ecológico-descritivo, transversal e de série temporal foi realizado em municípios do Maciço de Baturité, no estado de Ceará, com dados relativos à série histórica diária de casos e óbitos confirmados de indivíduos pelo novo coronavírus (covid-19) em municípios do Maciço de Baturité entre os dias 15/04/2020 e 17/09/2020.

Os municípios pesquisados foram de Acarape, Redenção, Aracoiaba e Baturité porque existe uma maior influência de estudantes e servidores da Universidade da Integração Internacional da Lusofonia Afro-Brasileira (2020) onde desenvolvemos as nossas atividades.

O banco de dados é formado por dados diários de casos e óbitos confirmados de indivíduos por coronavírus, baixados a partir do repositório de dados abertos sob licença Creative Commons (2020).

Nesse repositório, são compilados diariamente boletins epidemiológicos de 27 Secretarias Estaduais de Saúde (SES) que são disponibilizados em formato acessível à sociedade e foi utilizada neste estudo a tabela "caso_full", de 18 de setembro de 2020, que contempla um período observável entre 25/02/2020 e 17/09/2020, ou seja, 205 dias (sete meses). (BRASIL.IO, 2020).

De posse dos dados, foram feitos: a) limpeza; b) tratamento; c) análise e d) visualização dos dados. As informações do repositório se encontravam em formato aberto *comma-separated values* (CSV) (CREATIVYST SOFTWARE, 2020), o que facilita o acesso de dados pela população em geral.

Consideraram-se elegíveis para esta pesquisa todos os registros de boletins epidemiológicos diários relativos aos municípios de Maciço de Baturité no período de 25/02/2020 a 17/09/2020.

Este estudo se baseou no banco de dados, com um total de 729.072 casos confirmados e óbitos de indivíduos por coronavírus e 27.467 dados ausentes; isto corresponde a 3,8% do total do banco.

Este banco inclui todos os municípios brasileiros, considerando o período entre os dias 25/02/2020 e 17/09/2020. Apesar dessa baixa taxa de ausência, de 3,8%, o que não impactaria estatisticamente nos resultados

(BENNETT, 2001; SCHAEFER, 1999), optou-se por manter no estudo o banco de dados completo. Em grandes levantamentos epidemiológicos, como é o caso deste estudo, é quase inevitável a falta de dados – como, por exemplo, a falta de resposta do item, mas, após a coleta, a abordagem para lidar com dados ausentes nesses contextos é a imputação múltipla (RUBIN, 1987).

Contudo, aqui se utiliza da imputação das categorias mais frequentes estimadas por moda, usando a biblioteca Pandas, uma vez que esse demonstrou ser o método de imputações mais ajustado e, portanto, o que produz inferências mais seguras quando relacionado à aplicação de outras técnicas de imputação, tais como imputação por maior frequência usando Scikit-learn (categorias frequentes estimadas através de conjunto de treinamento) e imputação por maior frequência usando Feature-engine (categorias frequentes estimadas através de conjunto de treinamento) (GALLI, 2020).

Foram consideradas como variáveis de interesse para este estudo: i) *city* (nome de município); ii) *last_available_confirmed* (número de casos confirmados do último dia disponível igual ou anterior à data correspondente ao dia da coleta de dados); iii) *last_available_deaths* (número de mortes do último dia disponível igual ou anterior à data do dia da coleta de dados) e iv) *date* (data do dia de coleta dos dados).

As análises descritivas foram apresentadas por meio de tabelas de frequência e gráficos, com o uso de linguagem de programação Python (VAN ROSSUM; DRAKE, 2009), juntamente com o conjunto de bibliotecas estatísticas Pandas (MCKINNEY, 2010), NumPy (OLIPHANT, 2006), Matplotlib (HUNTER, 2007) e Plotly (PLOTLY, 2015).

A taxa de letalidade representa a proporção de casos de pessoas que eventualmente morrem de uma doença (óbitos/casos). Mas, enquanto uma epidemia ainda está em curso, como é o caso do novo surto de coronavírus, essa fórmula pode ser enganosa se, no momento da análise, o resultado for desconhecido para uma proporção não desprezível de pacientes (WORLDOMETERS, 2020).

Nesse sentido, usou-se, neste estudo, um método alternativo para o cálculo da taxa de letalidade por coronavírus [mortes/(mortes + recuperadas)], o mais ajustado (GHANI *et al*., 2005).

O presente estudo se utiliza de dados de domínio público e de livre acesso, sem a identificação dos participantes. Com isso, é desnecessária a apreciação no Sistema do Comitê de Ética em Pesquisa/Comissão Nacional de Ética em Pesquisa (CEP/Conep), em conformidade com as Resoluções nº 466, de 12 de dezembro de 2012; nº 510, de 7 de abril de 2016; e nº 580, de

Baturité é o [município] mais densamente povoado, proporcionalmente com menos casos confirmados, mas com uma maior taxa de letalidade pela covid-19 [...] Acarape é menos povoado, proporcionalmente tem mais casos confirmados, contudo, apresenta a menor taxa de letalidade por covid-19

22 de março de 2018, do Conselho Nacional de Saúde, que regulamentam as pesquisas com seres humanos e no âmbito do Sistema Único de Saúde, no Brasil (CONSELHO NACIONAL DE SAÚDE, 2012, 2016, 2018).

RESULTADOS

Os resultados obtidos indicam que no dia 15/04/2020 foram notificados os dois primeiros casos positivos de coronavírus em Redenção e um em Aracoiaba e um em Baturité. Por outro lado, após dois dias, i.e, no dia 18/04/2020, foram registrados os dois primeiros casos em Acarape, conforme é mostrado na Tabela 1.

Tabela 1
Dias de ocorrência dos primeiros registros de casos confirmados de coronavírus – Municípios do estado do Ceará

Cidade	Data	Casos confirmados
Fortaleza	16/03/2020	8
Redenção	15/04/2020	2
Aracoiaba	15/04/2020	1
Acarape	18/04/2020	2
Baturité	15/04/2020	1

Fonte: Elaborado pelos autores.

Entre os municípios em análise, Baturité é o mais densamente povoado, proporcionalmente com menos casos confirmados, mas com uma maior taxa de letalidade pela covid-19, de 5,24% no dia 17/09/2020. Por outro lado, Acarape é menos povoado, proporcionalmente tem mais casos confirmados, contudo, apresenta a menor taxa de letalidade por covid-19, de 0,83%, conforme é apresentado na Tabela 2.

Tabela 2
Distribuição de dados da covid-19 – Municípios do estado do Ceará – 17/09/2020

Cidade	População (1)	Casos confirmados	(%) população com coronavírus (covid-19)	Recuperados	Mortes	(%) Taxa de letalidade
Baturité	33.321	572	1,72	542	30	5,24
Redenção	26.415	1.420	5,38	1.382	38	2,68
Aracoiaba	25.391	763	3,01	749	14	1,83
Acarape	15.338	1.689	11,01	1.675	14	0,83

Fonte: Elaborado pelos autores.
(1) Censo Demográfico 2010 (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2011).

No município de Redenção, a taxa de letalidade varia entre 2,7% e 3,6%, a partir dos dados notificados 30 dias depois do registro de casos confirmados e durante 180 dias de observação. Porém, no dia 17/09/2020,

referente ao fim do período de análise, a taxa de letalidade no município de Redenção foi de aproximadamente 2,7%, o valor correspondente ao limite inferior da variação do período de análise em questão, conforme é apresentado na Tabela 3.

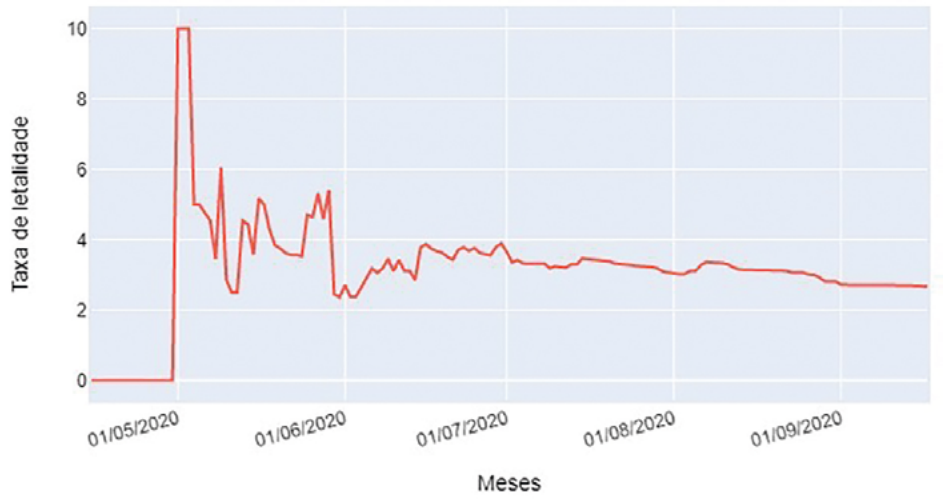
Tabela 3
Análise de séries temporais sobre covid-19 – Redenção (CE) – 15/04/2020–17/09/2020

Casos confirmados	Recuperados	Mortes	Taxa de mortalidade
Resultado da série no 12º dia de observação (Início da série: 15/04/2020 – Fim da série: 27/04/2020)			
9	9	0	0,00
Resultado da série no 30º dia de observação (Início da série: 15/04/2020 – Fim da série: 15/05/2020)			
56	54	2	3,57
Resultado da série no 60º dia de observação (Início da série: 15/04/2020 – Fim da série: 14/06/2020)			
491	477	14	2,85
Resultado da série no 90º dia de observação (Início da série: 15/04/2020 – Fim da série: 14/07/2020)			
970	938	32	3,30
Resultado da série no 120º dia de observação (Início da série: 15/04/2020 – Fim da série: 13/08/2020)			
1203	1165	38	3,16
Resultado da série no 150º dia de observação (Início da série: 15/04/2020 – Fim da série: 12/09/2020)			
1405	1367	38	2,70
Resultado da série no 180º dia de observação (Início da série: 15/04/2020 – Fim da série: 17/09/2020)			
1420	1382	38	2,68

Fonte: Elaborado pelos autores.

Para o período analisado, conforme é mostrado na Figura 1, a taxa de letalidade teve um pico no começo do período estudado e foi diminuindo para um valor de 2,7% em 17/09/2020. Porém, entre os dias 15 e 18, depois do início da observação de 01/05/2020 a 03/05/2020, ocorreram altas taxas de mortalidade, com valores atípicos, em torno de 10%.

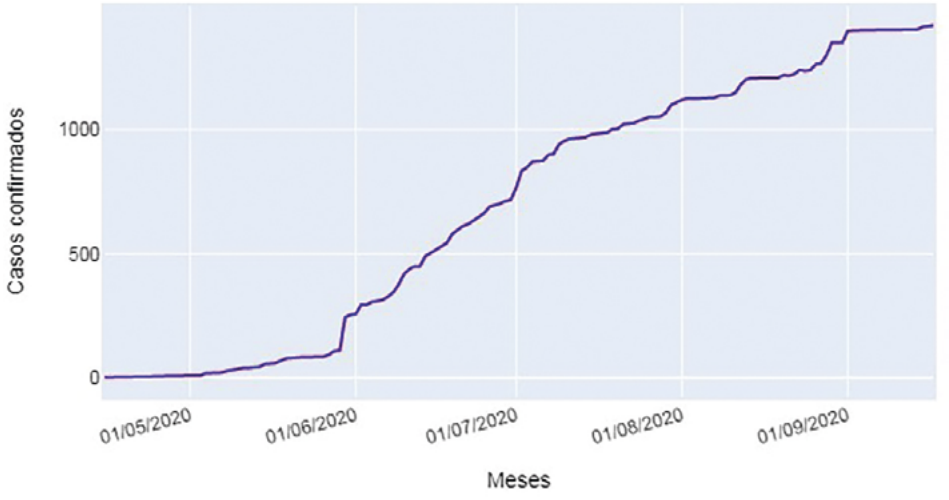
Figura 1
Taxa de letalidade pela covid-19 – Redenção (CE) – 15/04/2020-17/09/2020



Fonte: Elaborado pelos autores.

O número de casos confirmados em Redenção, conforme é mostrado na Figura 2, cresceu com o tempo e, a partir do dia 01/09/2020, apresenta estabilização em torno de 1.400 casos confirmados.

Figura 2
Casos confirmados pela covid-19 – Redenção (CE) – 15/04/2020-17/09/2020



Fonte: Elaborado pelos autores.

No município de Aracoiaba, a taxa de letalidade varia entre 0,0% e 1,9% a partir dos dados notificados em 30 dias depois do registro de casos confirmados e durante 180 dias de observação. Entretanto, no dia 17/09/2020, ao final do período de análise, a taxa de letalidade no município de Aracoiaba foi de, aproximadamente, 1,8%, conforme é apresentado na Tabela 4.

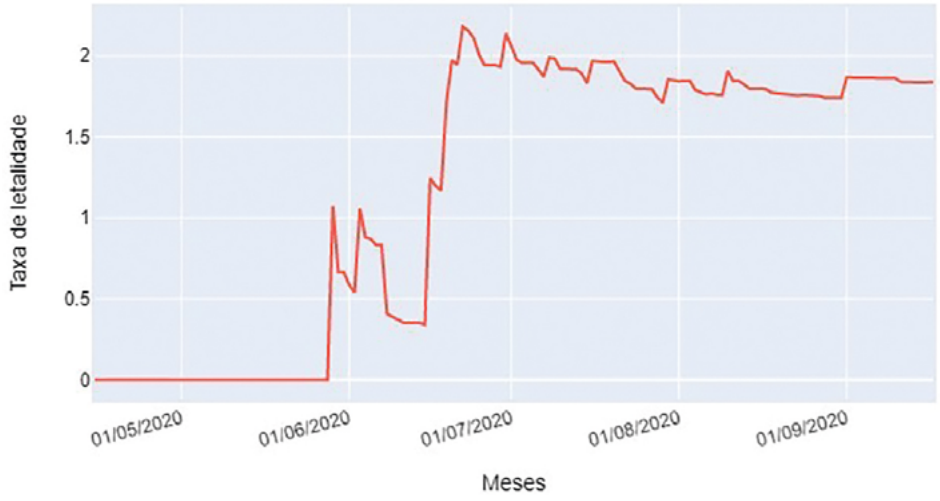
Tabela 4
Análise de séries temporais sobre a covid-19 – Aracoiaba (CE) – 15/04/2020-17/09/2020

Casos confirmados	Recuperados	Mortes	Taxa de mortalidade
Resultado da série no 12º dia de observação (Início da série: 15/04/2020 – Fim da série: 27/04/2020)			
7	7	0	0,00
Resultado da série no 30º dia de observação (Início da série: 15/04/2020 – Fim da série: 15/05/2020)			
45	45	0	0,00
Resultado da série no 60º dia de observação (Início da série: 15/04/2020 – Fim da série: 14/06/2020)			
283	282	1	0,35
Resultado da série no 90º dia de observação (Início da série: 15/04/2020 – Fim da série: 14/07/2020)			
529	519	10	1,89
Resultado da série no 120º dia de observação (Início da série: 15/04/2020 – Fim da série: 13/08/2020)			
713	700	13	1,82
Resultado da série no 150º dia de observação (Início da série: 15/04/2020 – Fim da série: 12/09/2020)			
762	748	14	1,84
Resultado da série no 180º dia de observação (Início da série: 15/04/2020 – Fim da série: 17/09/2020)			
762	748	14	1,84

Fonte: Elaborado pelos autores.

Há uma tendência de declínio de taxa de mortalidade no município de Aracoiaba, a partir do dia 22/06/2020, momento em que se vivenciou maior taxa de letalidade de 2,2% durante toda a série em análise, e, no dia 17/09/2020, a taxa ficou em torno de 1,8%, conforme é apresentado na Figura 3.

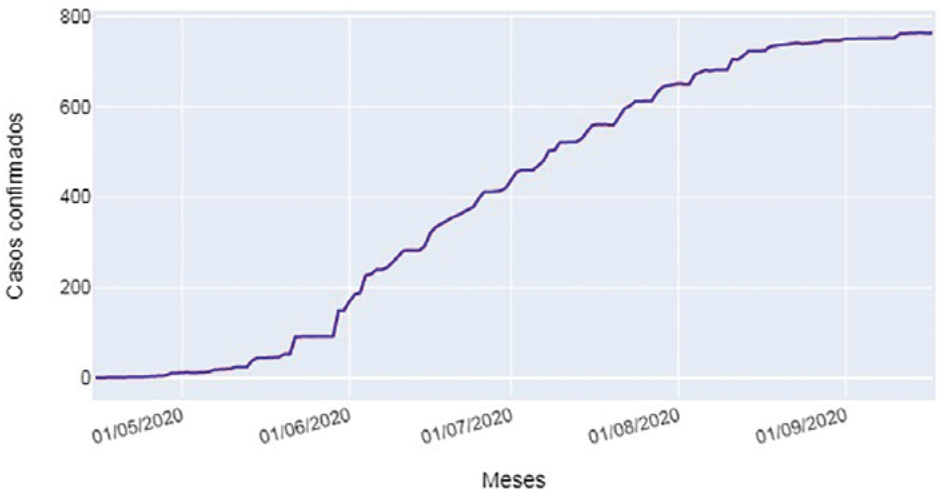
Figura 3
Taxa de letalidade pela covid-19 – Aracoiaba (CE) – 18/04/2020-17/09/2020



Fonte: Elaborado pelos autores.

No município de Aracoiaba, houve um crescimento mais lento no padrão da evolução de casos confirmados, conforme é apresentado na Figura 4, mostrando uma estabilidade em torno de 760 casos.

Figura 4
Casos confirmados pela covid-19 – Aracoiaba (CE) – 15/04/2020-17/09/2020



Fonte: Elaborado pelos autores.

No município de Acarape, a taxa de letalidade varia entre 0,8% e 15,4%, a partir dos dados notificados em 30 dias depois do registro de casos confirmados e durante 180 dias de observação. Por outro lado, no dia 17/09/2020, final de período de análise, a taxa de letalidade no município de Redenção foi de aproximadamente 0,8%, conforme é apresentado na Tabela 5.

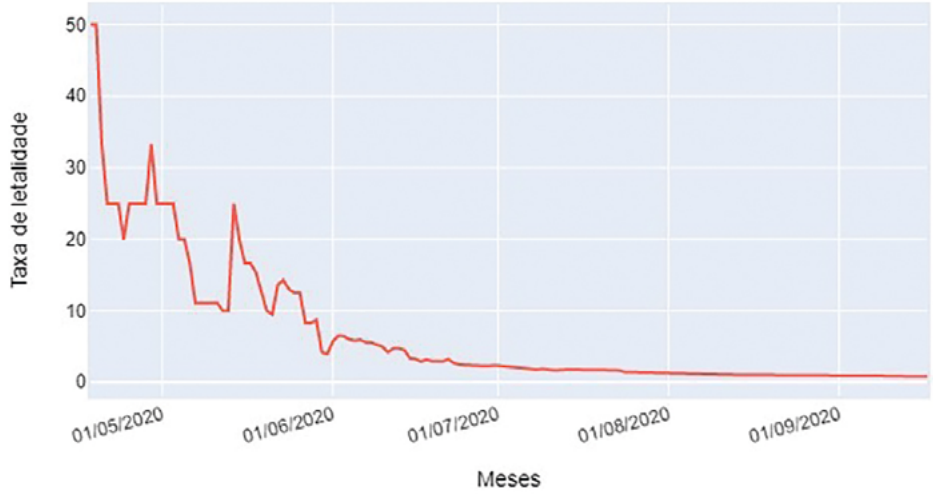
Tabela 5
Análise de séries temporais sobre covid-19 – Acarape (CE) – 15/04/2020-17/09/2020

Casos confirmados	Recuperados	Mortes	Taxa de mortalidade
Resultado da série no 12º dia de observação (Início da série: 18/04/2020 – Fim da série: 30/04/2020)			
4	3	1	25,00
Resultado da série no 30º dia de observação (Início da série: 18/04/2020 – Fim da série: 18/05/2020)			
13	11	2	15,38
Resultado da série no 60º dia de observação (Início da série: 18/04/2020 – Fim da série: 17/06/2020)			
310	301	9	2,90
Resultado da série no 90º dia de observação (Início da série: 18/04/2020 – Fim da série: 17/07/2020)			
807	793	14	1,73
Resultado da série no 120º dia de observação (Início da série: 18/04/2020 – Fim da série: 16/08/2020)			
1334	1320	14	1,05
Resultado da série no 150º dia de observação (Início da série: 18/04/2020 – Fim da série: 15/09/2020)			
1651	1637	14	0,85
Resultado da série no 180º dia de observação (Início da série: 18/04/2020 – Fim da série: 17/09/2020)			
1689	1675	14	0,83

Fonte: Elaborado pelos autores.

Em Acarape, houve uma taxa de letalidade elevada de 50%, no início do registro de casos confirmados de covid-19, entre os dias 17/04/2020 e 18/04/2020. Tal taxa foi diminuindo até atingir uma taxa de 0,83%, no dia 17/09/2020 (Figura 5).

Figura 5
Taxa de letalidade pela covid-19 – Acarape (CE) – 18/04/2020-17/09/2020

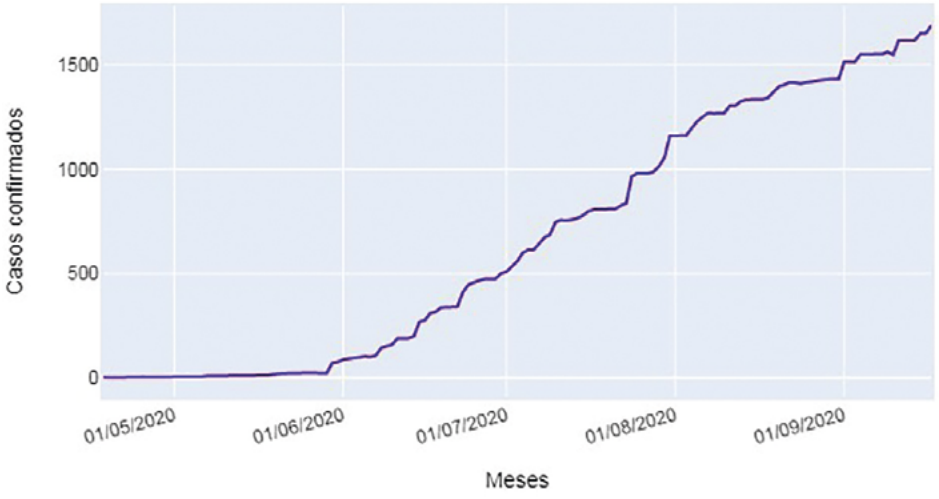


Fonte: Elaborado pelos autores.

O padrão da evolução de casos confirmados no município de Acarape apresenta um crescimento mais acelerado, conforme é mostrado na Figura 6. Este crescimento continua atingido a marca de cerca de 1.689 casos no dia 17/09/2020, último dia de observação.

O município de Baturité teve um índice de taxa de letalidade melhor que o de Acarape após os primeiros trinta dias de surgimento de casos confirmados do novo coronavírus

Figura 6
Casos confirmados pela covid-19 – Acarape (CE) – 15/04/2020-17/09/2020



Fonte: Elaborado pelos autores.

Em município de Baturité, a taxa de letalidade varia entre 5,2% e 11,8% (a partir dos dados notificados em 30 dias depois do registro de casos confirmados e durante 180 dias de observação). Entretanto, no dia 17/09/2020, o fim do período de análise, a taxa de letalidade no município de Redenção foi de aproximadamente 5,2% (Tabela 6). Baturité apresentou pior índice de taxa de letalidade que os de todos os outros municípios do Maciço de Baturité, em análise, ao longo de toda a série temporal em observação, e terminou com quase dois pontos percentuais a mais, no dia 17/09/2020, do que Redenção (5,2% e 2,7%, respectivamente, em Baturité e Redenção). O município de Baturité teve um índice de taxa de letalidade melhor que o de Acarape após os primeiros trinta dias de surgimento de casos confirmados do novo coronavírus (15,4% e 11,8, respectivamente, em Acarape e Baturité).

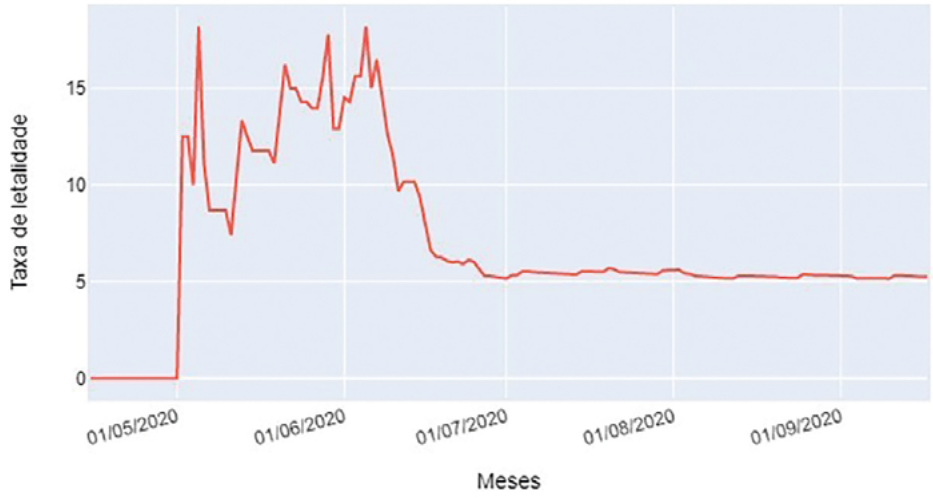
Em Baturité, parece haver estabilização da taxa de letalidade em torno de 5%, do dia 27/06/2020 até o dia 17/09/2020, o fim do período da série em análise, conforme é mostrado na Figura 7.

Tabela 6
Análise de séries temporais sobre covid-19 – Baturité (CE) – 15/04/2020-17/09/2020

Casos confirmados	Recuperados	Mortes	Taxa de mortalidade
Resultado da série no 12º dia de observação (Início da série: 15/04/2020 – Fim da série: 27/04/2020)			
4	4	0	0,00
Resultado da série no 30º dia de observação (Início da série: 15/04/2020 – Fim da série: 15/05/2020)			
34	30	4	11,76
Resultado da série no 60º dia de observação (Início da série: 15/04/2020 – Fim da série: 14/06/2020)			
187	168	19	10,16
Resultado da série no 90º dia de observação (Início da série: 15/04/2020 – Fim da série: 14/07/2020)			
466	441	25	5,36
Resultado da série no 120º dia de observação (Início da série: 15/04/2020 – Fim da série: 13/08/2020)			
529	501	28	5,29
Resultado da série no 150º dia de observação (Início da série: 15/04/2020 – Fim da série: 12/09/2020)			
563	533	30	5,33
Resultado da série no 180º dia de observação (Início da série: 15/04/2020 – Fim da série: 17/09/2020)			
572	542	30	5,24

Fonte: Elaborado pelos autores

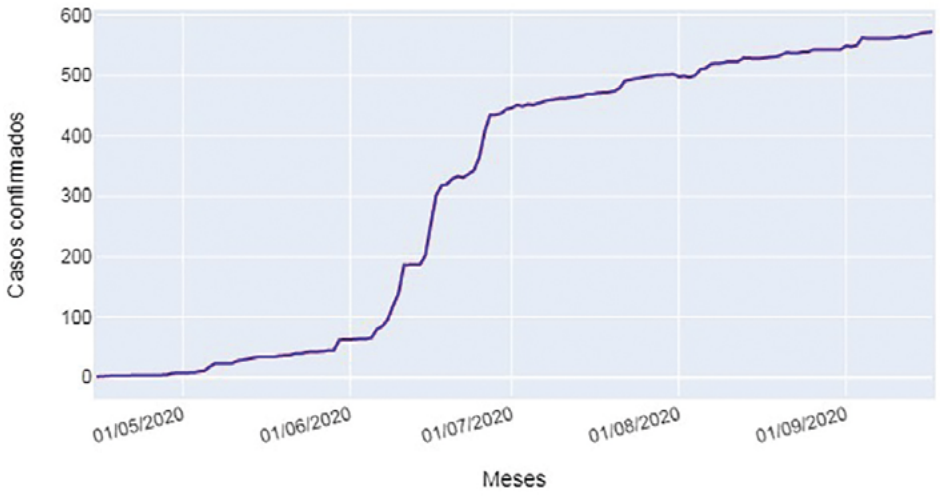
Figura 7
Taxa de letalidade pela covid-19 – Baturité (CE) – 15/04/2020-17/09/2020



Fonte: Elaborado pelos autores.

No município de Baturité, o padrão da evolução de casos confirmados teve um crescimento mais lento, variando entre 562 e 572, para os dias 04/09/2020 e 17/09/2020, respectivamente, conforme é apresentado na Figura 8.

Figura 8
Casos confirmados pela covid-19 – Baturité (CE) – 15/04/2020-17/09/2020



Fonte: Elaborado pelos autores.

Os municípios estudados neste trabalho apresentaram taxas semelhantes aos encontrados pela OMS e pelo Comitê Científico do Consórcio Nordeste

DISCUSSÃO

Ainda existem poucos artigos científicos de estudos populacionais, ecológicos e descritivos que exploram os dados de óbitos causados pelo novo coronavírus (covid-19) e casos confirmados fornecidos pelos boletins epidemiológicos de Secretarias Estaduais de Saúde (SES) no Brasil, em especial no Maciço de Baturité, no Ceará.

Em estudo anterior da Organização Mundial da Saúde (OMS) (WORLD HEALTH ORGANIZATION, 2020b), o organismo já afirmava que a taxa de letalidade por coronavírus no mundo era de 3,4%, valor considerado muito maior do que o que se esperava, de cerca de 2% (LOVELACE JR.; HIGGINS-DUNN, 2020).

No estudo realizado pelo Comitê Científico do Consórcio Nordeste (2020), através das principais fontes: Secretarias Estaduais de Saúde, Ministério da Saúde e Worldometers, a taxa de letalidade do coronavírus no Nordeste foi de 2,7%.

Os municípios estudados neste trabalho apresentaram taxas semelhantes aos encontrados pela OMS e pelo Comitê Científico do Consórcio Nordeste. Aracoiaba apresentou melhor índice de taxa de letalidade que Redenção ao longo de toda a série temporal em observação e terminou com quase um ponto percentual menor, no dia 17/09/2020, de 2,7% e 1,83% para Redenção e Aracoiaba, respectivamente. Estes valores estão próximos dos valores encontrados pelo Consórcio Nordeste de 2,4% e 1,88% para Redenção e Aracoiaba, respectivamente.

O padrão da evolução de casos confirmados em Acarape é diferente dos municípios de Redenção e Aracoiaba, pois apresentam sinais de estabilização de casos confirmado

Acarape apresentou melhor índice de taxa de letalidade que Aracoiaba e Redenção ao longo de toda a série temporal em observação e terminou com quase um ponto percentual a menos, no dia 17/09/2020, do que Aracoiaba, sendo de 1,8% para Aracoiaba e 0,8% para Acarape. Ressalta-se, que o município de Acarape teve um índice de taxa de letalidade pior que os de Aracoiaba e Redenção após os primeiros trinta dias de surgimento de casos confirmados de novo coronavírus, conforme apresentado nas Tabelas 3, 4 e 5.

Em comparação com os resultados obtidos pelo Consórcio Nordeste, de 0,88% para Acarape, se verifica uma grande semelhança com os resultados encontrados neste trabalho, de 0,8 para Acarape (COMITÊ CIENTÍFICO DO CONSÓRCIO NORDESTE, 2020).

Quando se trata do número de casos confirmados, o município de Acarape teve um crescimento mais acelerado, conforme é mostrado na Figura 6, em comparação à situação de Redenção, e até o dia 17/09/2020, último dia da série em análise, continua crescendo o número de casos confirmados, tendo atingido a marca de cerca de 1.689, neste último dia de observação. O padrão da evolução de casos confirmados em Acarape é diferente dos municípios de Redenção e Aracoiaba, pois apresentam sinais de estabilização de casos confirmado, conforme é apresentado nas Figuras 2 e 4.

No município de Baturité houve um crescimento mais lento de casos confirmados com o tempo, em comparação com os demais municípios analisados, indicando haver uma estabilização do número de casos confirmados, que variou entre 562 e 572, para os dias 04/09/2020 e 17/09/2020, respectivamente, conforme é apresentado na Figura 8.

Por outro lado, os resultados deste estudo mostram que, entre os municípios do Maciço de Baturité analisados, o município de Baturité, o mais densamente povoado, teve menos casos confirmados de coronavírus, entretanto, apresentou uma elevada taxa de letalidade por coronavírus (5,24%), maior que a taxa registrada em todo o Brasil (3%), no dia 17/09/2020 (ANSA BRASIL, 2020). Comparado com os resultados do Consórcio Nordeste, de 4,89% para Baturité, há uma variação de apenas 0,3 décimos de pontos percentuais, sendo estatisticamente pouco significativa (COMITÊ CIENTÍFICO DO CONSÓRCIO NORDESTE, 2020).

Os resultados deste trabalho revelam que a taxa de letalidade do município de Baturité também supera as dos estados de Ceará (3,9%), São Paulo (3,7%), Acre (2,4%), Rondônia (2,1%), Distrito Federal (1,7%) e Santa Catarina (1,3%) no dia 17/09/2020.

No mesmo sentido, essa taxa do município de Baturité é a maior entre os valores de taxa de letalidade por coronavírus. Inclusive, supera a da cidade chinesa de Wuhan na data de 4 de fevereiro (4,9%) (WORLDOMETER, 2020) e também ultrapassa em mais de dois pontos percentuais a taxa de mortalidade mundial, supracitada pela OMS (WORLD HEALTH ORGANIZATION, 2020b).

Achados deste estudo apontam que o estado do Rio de Janeiro e a cidade de Fortaleza apresentaram taxas de letalidade mais alta – 7,1% e 8,0%, respectivamente – quando comparados à situação dos municípios do Maciço de Baturité, em análise, embora pesquisas apontem que a taxa de letalidade por coronavírus é mais alta nas regiões mais pobres (BBC NEWS, 2020; FORTALEZA, 2020; JANSEN, 2020; LEÃO *et al*., 2020; VIANA, 2020), não obstante os primeiros casos terem surgido em regiões mais ricas.

Após quase cinco meses de pandemia, tais regiões já concentravam as maiores taxas de letalidade por coronavírus, sendo até 18 vezes maiores que as da área nobre da capital [Fortaleza]

Nesse sentido, no município de Fortaleza, a principal cidade do estado do Ceará, os primeiros casos confirmados de coronavírus surgiram nos bairros Meireles e Aldeota, que estão entre aqueles com Índice de Desenvolvimento Humano (IDH) muito alto, reflexo ainda do início da epidemia. Aos poucos, com a circulação dos indivíduos que residem em regiões mais pobres e trabalham nas ricas, a enfermidade migrou para as periferias. Após quase cinco meses de pandemia, tais regiões já concentravam as maiores taxas de letalidade por coronavírus, sendo até 18 vezes maiores que as da área nobre da capital (FORTALEZA, 2020).

Sendo assim, atualmente, no município de Fortaleza, a Regional II, que inclui Meireles e Aldeota, é a que possui mais casos confirmados de coronavírus. Entretanto, é a Regional V, de bairros como Conjunto Ceará e Bom Jardim, que concentra as maiores taxas de letalidade por covid-19 (FORTALEZA, 2020).

Tendência semelhante está ocorrendo em outros estados, como, por exemplo, São Paulo (FUNDAÇÃO SISTEMA ESTADUAL DE ANÁLISE DE DADOS, 2020), onde houve um aumento de 45% nas mortes nos 20 distritos mais pobres da cidade (VESPA, 2020).

A experiência europeia – como mostra, por exemplo, um estudo realizado na Inglaterra (BBC NEWS, 2020) – também tem indicado que indivíduos mais carentes da Inglaterra e do País de Gales têm maior probabilidade de morrer vítima de coronavírus do que aqueles de lugares mais ricos, conforme sugerem novos números. A análise do Office for National Statistics revela que ocorreram 55,1 óbitos para cada 100 mil indivíduos nas regiões mais pobres da Inglaterra, em comparação com as 25,3 mortes registradas para cada 100 mil pessoas nas áreas mais

ricas (CAUL, 2020). A taxa de mortalidade é também mais elevada entre os homens, com 76,7 óbitos a cada 100 mil pessoas, em comparação com 39,6 por 100 mil mulheres (CAUL, 2020).

A diferença na taxa de letalidade entre homens e mulheres ingleses pode ser melhor compreendida quando analisada como uma questão de gênero, e não de sexo biológico, em que os homens são incentivados a se mostrarem como um ser mais valente, ou seja, que não tem medo da doença, enquanto as mulheres seriam mais frágeis e desencorajadas a esse enfrentamento. Infelizmente, neste estudo não é possível se testar a hipótese desse diferencial, por conta de ausência da variável "sexo" no banco de dados secundários utilizado.

A diferença na taxa de letalidade por coronavírus entre regiões mais pobres e ricas, especialmente no Brasil, pode ser entendida como uma questão de estrutura, e não de responsabilização de indivíduos, uma vez que existe um número importante de indivíduos que não têm condições de cumprir o isolamento

Provavelmente, a diferença na taxa de letalidade por coronavírus entre regiões mais pobres e ricas, especialmente no Brasil, pode ser entendida como uma questão de estrutura, e não de responsabilização de indivíduos, uma vez que existe um número importante de indivíduos que não têm condições de cumprir o isolamento por morarem em casas muito pequenas, onde há vários indivíduos dividindo um mesmo quarto, o que pode revelar como a sociedade brasileira ainda está atrasada quando se trata de condições dignas de habitação e saneamento básico (LUPION, 2020).

O isolamento social tem como principal objetivo reduzir as interações entre indivíduos de uma comunidade, na qual estão contidos infectados confirmados e aqueles não identificados que não se encontravam isolados. Esse tipo de ingerência é necessário para o achatar a curva epidemiológica, impedindo a transmissão desregrada e a superlotação do sistema de saúde, que ganhará tempo para se estruturar (GARCIA; DUARTE, 2020; LIMA *et al*., 2020). Todavia, para ser realizar-se, essa medida depende de vários fatores, principalmente, sociais, culturais e econômicos (LIMA *et al*., 2020).

Nessa direção, o conhecimento, a identificação e análise de padrões da evolução de taxa de letalidade por coronavírus e casos confirmados são decisivos para determinar o ajustamento das estratégias de mitigação e para permitir o planejamento das necessidades de cuidados de saúde conforme a desenvoltura da epidemia. No entanto, as razões brutas de mortalidade obtidas pela divisão do número de mortes pelo número de casos podem ser enganosas (GARSKE *et al*., 2009; LIPSITCH *et al*., 2015). Primeiro, pode haver um período de duas a três semanas entre o desenvolvimento dos sintomas de uma pessoa, a detecção e notificação do caso subsequente e a observação do resultado clínico final (GARSKE *et al*., 2009; VERITY *et al*., 2020).

Portanto, muitas das estimativas da razão de letalidade que foram obtidas até o momento para covid-19 corrigem esse efeito (VERITY *et*

al., 2020). As estimativas da razão de letalidade podem, portanto, ser enviesadas para cima até que a extensão da doença clinicamente mais branda seja determinada (VERITY *et al*., 2020).

Sendo assim, neste estudo, usou-se um método alternativo, mais robusto para o cálculo da taxa de letalidade por coronavírus (mortes / (mortes + recuperadas), segundo (GHANI *et al*., 2005).

CONCLUSÃO

Os resultados obtidos apontam que a pandemia se desenvolve em diversas fases e, neste momento, os surtos localizados foram superados, adentrando a fase de estabilização/desaceleração em todos os municípios analisados, com exceção ao de Acarape, o único município sem sinal de estabilização dos casos confirmados até o dia 17/09/2020 e último a registrar a ocorrência de casos confirmados de coronavírus.

Além disso, o município menos densamente povoado e com a menor taxa de letalidade por coronavírus de 0,83%, no dia 17/09/2020. Por outro lado, Baturité é o município mais densamente povoado, com menos casos confirmados; entretanto, com uma maior taxa de letalidade de 5,24% por covid-19.

O conhecimento, a identificação e análise de padrões da evolução de taxa de letalidade por coronavírus e casos confirmados são decisivos para determinar o ajustamento das estratégias de mitigação e para permitir o planejamento das necessidades de cuidados de saúde conforme a desenvoltura da epidemia.

O tema desta edição da revista *Bahia Análise e Dados* gera possibilidades dialógicas importantes entre cientistas, poderes públicos, entes privados e a sociedade como um todo, reforçando as diversas aberturas ou pontes para se trabalhar com grandes volumes de dados em prol de uma gestão da coisa pública comprometida com o bem-estar e a saúde dos cidadãos e cidadãs.

Quando governos, de todas as esferas, orientam-se por estratégias técnico-científicas e metodologias consistentes em suas ações e políticas públicas capazes de gerar benefícios sociais a todos e todas sem distinção, a ciência cumpre um dos seus mais relevantes papéis; e esperamos que isso ocorra, com maior contundência, com relação à pandemia que vivenciamos.

Afinal, até a população inteira ser vacinada, quando forem disponibilizadas vacinas efetivamente eficazes após o período analisado neste estu-

Quando governos, de todas as esferas, orientam-se por estratégias técnico-científicas e metodologias consistentes em suas ações e políticas públicas capazes de gerar benefícios sociais a todos e todas sem distinção, a ciência cumpre um dos seus mais relevantes papéis

do, haverá ainda um considerável tempo; e, paralelamente a isso, a missão governamental e científica de prover investimentos e insumos para lidarmos com pandemias, como esta da covid-19, e até mesmo para preveni-las, deve ser cobrada e alertada, em todo momento, diante do atual panorama permeado de possibilidades mas, também, de incertezas.

Aproveitamos para agradecer à Superintendência de Estudos Econômicos e Sociais da Bahia o honroso convite, colocando-nos à disposição para mais parcerias futuras.

REFERÊNCIAS

ALAGOAS. Secretaria de Estado do Planejamento, Gestão e Patrimônio. *Painel Covid-19 em Alagoas*. Disponível em: https://dados.al.gov.br/catalogo/dataset/painel-covid-19-em-alagoas. Acesso em: 19 out. 2020.

ANSA BRASIL. *Brasil tem 829 mortes e 36,3 mil novos casos de Covid em 24h*. São Paulo, 18 set. 2020. Disponível em: http://ansabrasil.com.br/brasil/noticias/americalatina/brasil/2020/09/17/brasil-tem-829-mortes-e-363-mil-novos-casos-de-covid-em-24h_08bfe0a6-9e98-4b1a-8cfb-434b87dab433.html. Acesso em: 15 out. 2020.

AQUINO, E. M. L. *et al*. Medidas de distanciamento social no controle da pandemia de COVID-19: potenciais impactos e desafios no Brasil. *Ciência & Saúde Coletiva*, Rio de Janeiro, v. 25, p. 2423-2446, jun. 2020. Supl.1. Disponível em: https://www.scielo.br/pdf/csc/v25s1/1413-8123-csc-25-s1-2423.pdf. Acesso em: 19 out. 2020.

BBC NEWS. *Coronavirus*: higher death rate in poorer areas, ONS figures suggest. United Kingdom, 1 May 2020. Disponível em: https://www.bbc.com/news/uk-52506979. Acesso em: 16 out. 2020.

BENNETT, D. A. How can I deal with missing data in my study?. *Australian and New Zealand Journal of Public Health*, [s. l.], v. 25, n. 5, p. 464-9, Oct. 2001. Disponível em: https://pubmed.ncbi.nlm.nih.gov/11688629/. Acesso em: 3 dez. 2020.

BOLETIM EPIDEMIOLÓGICO. Brasília: MS, n. 3, 21 fev. 2020. Disponível em: https://portalarquivos.saude.gov.br/images/pdf/2020/fevereiro/21/2020-02-21-Boletim-Epidemiologico03.pdf . Acesso em: 18 out. 2020.

BRASIL. Ministério da Saúde. *Coronavírus (COVID-19)*: o que você precisa saber. Disponível em: https://coronavirus.saude.gov.br. Acesso em: 18 out. 2020a.

BRASIL. Ministério da Saúde. Secretaria de Atenção Primária à Saúde. *Protocolo de manejo clínico do coronavírus (COVID-19) na atenção primária à saúde*. Versão 9. Brasília: MS, maio 2020b. Disponível em: https://www.unasus.gov.br/especial/covid19/pdf/37. Acesso em: 19 out. 2020.

BRASIL.IO. *COVID-19*: boletins informativos e casos do coronavírus por município por dia. Disponível em: https://www.brasil.io/dataset/covid19/caso_full/. Acesso em: 18 set. 2020.

CAUL, S. *Deaths involving COVID-19 by local area and socioeconomic deprivation*: deaths occurring between 1 March and 17 April 2020. [S. l.]: ONS, 2020. 23 p. Disponível em: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19bylocalareasanddeprivation/deathsoccurringbetween1marchand17april. Acesso em: 16 out. 2020.

CREATIVE COMMONS. CEARÁ. Decreto nº 33.544, de 19 de abril de 2020. Prorroga, em âmbito estadual, as medidas necessárias ao enfrentamento da pandemia da COVID-19, e dá outras providências. *Diário Oficial [do] Estado do Ceará*, Fortaleza, 19 abr. 2020a. Disponível em: https://www.ceara.gov.br/wp-content/uploads/2020/04/DECRETO-N%C2%BA33.544-de-19-de-abril-de-2020.pdf. Acesso em: 9 out. 2020.

CEARÁ. Secretaria da Saúde. *Ceará confirma três casos do novo coronavírus*. Fortaleza, 15 mar. 2020b. Disponível em: https://www.saude.ce.gov.br/2020/03/15/ceara-confirma-tres-casos-do-novo-coronavirus/. Acesso em: 9 out. 2020.

COMITÊ CIENTÍFICO DO CONSÓRCIO NORDESTE. *Painéis de dados*. Disponível em: https://www.comitecientifico-ne.com.br/. Acesso em: 1 dez. 2020.

CONSELHO NACIONAL DE SAÚDE. Resolução nº 466, de 12 de dezembro de 2012. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 12 dez. 2012. Disponível em: https://bvsms.saude.gov.br/bvs/saudelegis/cns/2013/res0466_12_12_2012.html. Acesso em: 9 out. 2020.

CONSELHO NACIONAL DE SAÚDE. Resolução nº 510, de 7 de abril de 2016. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 7 abr. 2016. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/cns/2016/res0510_07_04_2016.html. Acesso em: 9 out. 2020.

CONSELHO NACIONAL DE SAÚDE. Resolução nº 580, de 22 de março de 2018. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 16 jul. 2018. Disponível em: https://conselho.saude.gov.br/resolucoes/2018/Reso580.pdf. Acesso em: 9 out. 2020.

CREATIVYST SOFTWARE. *Creativyst Docs*: understanding CSV file formats. Disponível em: http://creativyst.com/Doc/Articles/CSV/CSV01.htm. Acesso em: 19 out. 2020.

CRODA, J. *et al*. COVID-19 in Brazil: advantages of a socialized unified health system and preparation to contain cases. *Revista da Sociedade Brasileira de Medicina Tropical*, Uberaba, v. 53, p. 1-6, 2020. Disponível em: https://www.scielo.br/pdf/rsbmt/v53/1678-9849-rsbmt-53-e20200167.pdf. Acesso em: 19 out. 2020.

FORTALEZA. Secretaria Municipal da Saúde. *Informe semanal COVID-19: 22ª semana epidemiológica*. Fortaleza: SMS, 2020. Disponível em: https://coronavirus.fortaleza.ce.gov.br/pdfs/informe-semanal-covid-19-se-22a-2020-sms-fortaleza.pdf . Acesso em: 16 out. 2020.

FUNDAÇÃO SISTEMA ESTADUAL DE ANÁLISE DE DADOS. *SP contra o novo coronavírus*: boletim completo. São Paulo: SEADE, 2020. Disponível em: https://www.seade.gov.br/coronavirus/. Acesso em: 16 out. 2020.

GALLI, S. *Python feature engineering cookbook*. Birmingham: Packt Publishing, 2020. v. 1.

GALLI, S. *Python Feature Engineering Cookbook*. 1. ed. B372 p.

GARCIA, L. P.; DUARTE, E. Intervenções não farmacológicas para o enfrentamento à epidemia da COVID-19 no Brasil. *Epidemiologia e Serviços de Saúde*, Brasília, v. 29, n. 2, p. 1-4, 2020. Disponível em: https://www.scielo.br/pdf/ress/v29n2/2237-9622-ress-29-02-e2020222.pdf. Acesso em: 20 out. 2020.

GARSKE, T. *et al*. *Assessing the severity of the novel influenza A/H1N1 pandemic*. London: BMJ, 14 July 2009. (BMJ, 339). Disponível em: http://www.bmj.com/content/339/bmj.b2840.abstract. Acesso em: 9 out. 2020.

GHANI, A. C. *et al*. Methods for estimating the case fatality ratio for a novel, emerging infectious disease. *American Journal of Epidemiology*, [s. l.], v. 162, n. 5, p. 479–486, Sept. 2005. Disponível em: https://doi.org/10.1093/aje/kwi230. Acesso em: 15 out. 2020.

HUNTER, J. D. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering*, [s. l.], v. 9, n. 3, p. 90-95, 2007.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Sinopse do censo demográfico 2010*. Rio de Janeiro: IBGE, 2011. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv49230.pdf. Acesso em: 20 out. 2020.

ITALIA. Ministero della Salute. Dipartimento della Protezione Civile. *COVID-19 Situazione Italia*. Disponível em: http://opendatadpc.maps.arcgis.com/apps/opsdashboard/index.html#/b0c68bce2cce478eaac82fe38d4138b1. Acesso em: 19 out. 2020.

JANSEN, R. Desigualdade leva covid-19 a matar mais nos bairros pobres do Rio. *O Estado de São Paulo*, São Paulo, 4 ago. 2020. Disponível em: https://saude.estadao.com.br/noticias/geral,desigualdade-leva-covid-19-a-matar-mais-nos-bairros-pobres-do-rio,70003386714. Acesso em: 16 out 2020.

LANA, R. M. *et al*. Emergência do novo coronavírus (SARS-CoV-2) e o papel de uma vigilância nacional em saúde oportuna e efetiva. *Cadernos de Saúde Pública*, Rio de Janeiro, v. 36, n. 3, 2020. Disponível em: https://www.scielo.br/pdf/csp/v36n3/1678-4464-csp-36-03-e00019620.pdf. Acesso em: 12 out. 2020.

LEÃO, A. L. *et al*. Covid-19 é mais letal em regiões de periferia no Brasil. *O Globo*, São Paulo, 3 maio 2020. Disponível em: https://oglobo.globo.com/sociedade/covid-19-mais-letal-em-regioes-de-periferia-no-brasil-1-24407520. Acesso em: 16 out. 2020.

LIMA, C. M. A. O. Informações sobre o novo coronavírus (COVID-19). *Radiologia Brasileira*, São Paulo, v. 53, n. 2, p. 1-2, mar./abr. 2020. Disponível em: https://www.scielo.br/pdf/rb/v53n2/pt_0100-3984-rb-53-02-000V.pdf. Acesso em: 19 out. 2020.

LIMA, D. L. F. *et al*. COVID-19 no estado do Ceará, Brasil: comportamentos e crenças na chegada da pandemia. *Ciência & Saúde Coletiva*, Rio de Janeiro, v. 25, n. 5, p. 1575-1586, 2020. Disponível em: https://www.scielo.br/pdf/csc/v25n5/1413-8123-csc-25-05-1575.pdf. Acesso em: 20 out. 2020.

LIMA-COSTA, M. F.; BARRETO, S. M. Tipos de estudos epidemiológicos: conceitos básicos e aplicações na área do envelhecimento. *Epidemiologia e Serviços de Saúde*, Brasília, v. 12, n. 4, p. 189-201, dez. 2020. Disponível em: http://scielo.iec.gov.br/pdf/ess/v12n4/v12n4a03.pdf. Acesso em: 19 out. 2020.

LIPSITCH, M. *et al*. Potential biases in estimating absolute and relative case-fatality risks during outbreaks. *PLOS Neglected Tropical Diseases*, United States, 16 July 2015. Disponível em: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0003846. Acesso em: 9 out. 2020.

LOVELACE JR., B.; HIGGINS-DUNN, N. *WHO says coronavirus death rate is 3.4% globally*, higher than previously thought. United States, 3 Mar. 2020. Disponível em: https://www.cnbc.com/2020/03/03/who-says-coronavirus-death-rate-is-3point4percent-globally-higher-than-previously-thought.html. Acesso em: 9 out. 2020.

LUPION, B. *Como o novo coronavírus acentua as desigualdades no Brasil*. Brasil, 27 abr. 2020. Disponível em: https://www.dw.com/pt-br/como-o-novo-corona-v%C3%ADrus-acentua-as-desigualdades-no-brasil/a-53256164. Acesso em: 19 out. 2020.

MADEIRO, C. CE e MA apontam benefícios do lockdown e iniciam reabertura da economia. *UOL Notícias*, Maceió, 31 maio 2020. Disponível em: https://noticias.uol.com.br/saude/ultimas-noticias/redacao/2020/05/31/ce-e-ma-freiam-covid-19-apos-lockdown-e-iniciam-reabertura-da-economia.htm. Acesso em: 19 out. 2020.

MARINELLI, N. P. *et al*. Evolução de indicadores e capacidade de atendimento no início da epidemia de COVID-19 no Nordeste do Brasil, 2020. *Epidemiologia e Serviços de Saúde*, Brasília, v. 29, n. 3, 2020. Disponível em: https://www.scielo.br/pdf/ress/v29n3/2237-9622-ress-29-03-e2020226.pdf. Acesso em: 19 out. 2020.

MCKINNEY, W. Data structures for statistical computing in Python. In: PYTHON IN SCIENCE CONFERENCE, 9., 2010, Austin, TX. *Proceedings* [...]. Austin, TX: SCIPY, 2010. p. 56-61. Disponível em: http://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf. Acesso em: 9 out. 2020.

MORAES, B. Q. S. *et al*. Análise dos indicadores da COVID-19 no Nordeste brasileiro em quatro meses de pandemia. *Vigilância Sanitária em Debate*: Sociedade, Ciência & Tecnologia, Rio de Janeiro, v. 8, n. 3, p. 52-60, 2020. Disponível em: https://visaemdebate.incqs.fiocruz.br/index.php/visaemdebate/article/view/1690/1200. Acesso em: 19 out. 2020.

NUNES, M. D. R. *et al*. Exames diagnósticos e manifestações clínicas da Covid-19 em crianças: revisão integrativa. *Texto & Contexto* - Enfermagem, Florianópolis, v. 29, 2020. Disponível em: https://www.scielo.br/pdf/tce/v29/pt_1980-265X-tce-29-e20200156.pdf. Acesso em: 18 out. 2020.

NYTIMES. The New York TimesOLIPHANT, T. E. *Guide to NumPy*. 2006. Disponível em: https://web.mit.edu/dvp/Public/numpybook.pdf. Acesso em: 9 out. 2020.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. *Folha informativa COVID-19 - Escritório da OPAS e da OMS no Brasil*. Disponível em: https://www.paho.org/pt/covid19. Acesso em: 18 out. 2020.

PIRES, L. S. *et al*. O mapa do coronavírus: como aumentam os casos dia a dia no Brasil e no mundo. *El País*, [s. l.], 20 nov. 2020. Disponível em: https://brasil.elpais.com/brasil/2020/03/12/ciencia/1584026924_318538.html . Acesso em: 14 out. 2020.

PLOTLY. Plotly Technologies Inc. Plotly Python Open Source Graphing Library. Collaborative data science. 2015. Disponível em: https://plotly.com/graphing-libraries/ . Acesso em: 9 out. 2020.

PRADO, M. F. *et al*. Analysis of COVID-19 under-reporting in Brazil. *Revista Brasileira de Terapia Intensiva*, São Paulo, v. 32, n. 2, p. 224-228, 2020. Disponível em: https://www.scielo.br/pdf/rbti/v32n2/0103-507X-rbti-20200030.pdf. Acesso em: 19 out. 2020.

RAWORTH, K. *Doughnut economics*: seven ways to think like a 21st-century economist. Chelsea: Chelsea Green Publishing, 2017. 384 p. Disponível em: https://www.kateraworth.com/. Acesso em: 19 out. 2020.

RECOMMENDATIONS for the diagnosis, prevention and control of the 2019 novel coronavirus infection in children. *Chinese Journal of Pediatrics*, Wuhan, v. 58, n. 3, p. 169-174, Mar. 2020. Disponível em: http://rs.yiigle.com/CN112140202003/1183499.htm. Acesso em: 18 out. 2020.

RIBEIRO, V. *Brasil registra 5,2 milhões de casos de covid-19 com 153.214 mortes*. Brasília, 16 out. 2020. Disponível em: https://agenciabrasil.ebc.com.br/radioagencia-nacional/saude/audio/2020-10/brasil-registra-52-milhoes-de-casos-de-covid-19-com-153214-mortes. Acesso em: 19 out. 2020.

RODRIGUES, A. *Ministério da Saúde confirma primeiro caso de coronavírus no Brasil*. Brasília, 26 fev. 2020. Disponível em: https://agenciabrasil.ebc.com.br/saude/noticia/2020-02/ministerio-da-saude-confirma-primeiro-caso-de-coronavirus-no-brasil. Acesso em: 9 out. 2020.

RUBIN, D. B. *Multiple imputation for nonresponse* in surveys. New York: John Wiley & Sons, 1987. 258 p.

SCHAEFER, L. Multiple imputation: a primer. *Statistical Methods in Medical Research*, [s. l.], v. 8, n. 1, p. 3-15, Mar 1999. Disponível em: https://pubmed.ncbi.nlm.nih.gov/10347857/. Acesso em: 3 dez. 2020.

SILVA, C. M.; ARBILLA, G. COVID-19: challenges for a new epoch. *Revista da Sociedade Brasileira de Medicina Tropical*, Uberaba, v. 53, 2020. Disponível em: https://www.scielo.br/pdf/rsbmt/v53/1678-9849-rsbmt-53-e20200270.pdf. Acesso em: 19 out. 2020.

SMSSOCIETY OF PEDIATRICS; CHINESE MEDICAL ASSOCIATION; EDITORIAL BOARD; CHINESE JOURNAL OF PEDIATRICS. RecommendationsTHE NEW YORK TIMES. *Coronavirus world map*: tracking the global outbreak. Disponível em: https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html. Acesso em: 18 out. 2020.

UNIVERSIDADE DA INTEGRAÇÃO INTERNACIONAL DA LUSOFONIA AFRO-BRASILEIRA. *Unilab em números*: ações realizadas. 2020. Disponível em: https://app.powerbi.com/view?r=eyJrIjoiOGZjMDBlY2MtNDM1OC00NDAwLWIzNDEtMGJkNjQ5NDkxMzRlIiwidCI6IjkwMjlkZGNlLWFmMTItNDJiZS04MDM3LTU4MzEzZTRkYzVkMSJ9. Acesso em: 2 dez. 2020.

VAN ROSSUM, G.; DRAKE, F. L. *Python 3*: a referência da linguagem Python. 2009. Disponível em: https://docs.python.org/pt-br/3/reference/. Acesso em: 9 out. 2020.

VERITY, R. *et al*. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*, [s. l.], v. 20, n. 6, p. 669-677, 2020. Disponível em: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30243-7/fulltext. Acesso em: 9 out. 2020.

VESPA, T. *Em vez da idade, classe social passa a definir quem morre de covid no país*. São Paulo, 6 maio 2020. Disponível em: https://noticias.uol.com.br/saude/ultimas-noticias/redacao/2020/05/06/no-brasil-covid-19-nao-mata-por-idade-mas-por-endereco-sugere-estudo.htm. Acesso em: 16 out. 2020.

VIANA, T. Covid-19: taxa de letalidade é até 18 vezes mais alta na periferia. *Diário do Nordeste*, Fortaleza, 1 ago. 2020. Disponível em: https://diariodonordeste.verdesmares.com.br/metro/covid-19-taxa-de-letalidade-e-ate-18-vezes-mais-alta-na-periferia-1.2972563. Acesso em: 16 out. 2020.

VILLELA, D. A. M. The value of mitigating epidemic peaks of COVID-19 for more effective public health responses. *Revista da Sociedade Brasileira de Medicina Tropical*, Uberaba, v. 53, 2020. Disponível em: https://www.scielo.br/pdf/rsbmt/v53/1678-9849-rsbmt-53-e20200135.pdf. Acesso em: 19 out. 2020.

WORLD HEALTH ORGANIZATION. *Coronavirus disease (COVID-19) dashboard*. Disponível em: https://covid19.who.int/region/wpro/country/cn . Acesso em: 19 out. 2020a.

WORLD HEALTH ORGANIZATION. *Coronavirus disease (COVID-19) pandemic*. Disponível em: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Acesso em: 9 out. 2020b.

WORLD HEALTH ORGANIZATION. *Prioritizing diseases for research and development in emergency contexts*. Disponível em: https://www.who.int/activities/prioritizing-diseases-for-research-and-development-in-emergency-contexts. Acesso em: 12 out. 2020c.

WORLD HEALTH ORGANIZATION. *Strengthening health security by implementing the International Health Regulations (2005)*. Disponível em: https://www.who.int/ihr/procedures/pheic/en/. Acesso em: 14 out. 2020d.

WORLDOMETERS. *Coronavirus (COVID-19) mortality rate*. Disponível em: https://www.worldometers.info/coronavirus/coronavirus-death-rate/. Acesso em: 15 out. 2020.

WU, A. *et al*. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host & Microbe*, [s. l.], v. 27, n. 3, p. 325-328, Mar. 2020. Disponível em: https://reader.elsevier.com/reader/sd/pii/S1931312820300727?token=9A35FBDB0F77646E033D8CBA910AFF77806C65463FD3C63AE4E7C1128B6A0E36D1F5724BF4F297E89BDC4E3B3A2FF1C2. Acesso em: 18 out. 2020.

Resumo

O Programa Bolsa Família impacta positivamente diversos aspectos da sociedade, desde a melhora nos indicadores das áreas da educação e saúde até impactos econômicos nas famílias beneficiadas e nos municípios. O Portal da Transparência disponibiliza um volume mensal de dados dos pagamentos do Bolsa Família, os quais totalizam cerca de 15 milhões de pagamentos ao mês. Assim, faz-se necessária a utilização de ferramentas que permitam a exploração e o entendimento de um grande volume de dados, bem como o emprego de ferramentas de análise de dados que fomentem boas práticas na gestão pública. O presente artigo utiliza-se da agregação de dados nacionais acerca da utilização do benefício do Programa Bolsa Família em um recorte para os municípios baianos. Através do processamento de mais de 100 GB de dados públicos via Apache Spark e Linguagem R, conjuntamente com a utilização do método de Máquinas Aleatórias, os resultados exibem importantes variáveis explicativas relacionadas à predição da taxa de utilização municipal do Bolsa Família no estado da Bahia.
Palavras-chave: Bolsa Família. Spark. R. Big Data.

Abstract

The Bolsa Família Program has a positive impact on several aspects of society, from the improvement on indicators in the areas of education, health to economic impacts on benefit of families and municipalities. The Transparency Portal provides a monthly volume of Bolsa Família payments data, which totalize approximately 15 million monthly payments. Thus, it is necessary to use tools that allow the exploration and understanding of a large volume of data, as well as the use of data analysis tools that encourages good practices in public management. This paper uses the aggregation of national data on the use of the Bolsa Família Program benefit in Bahia's municipalities profile. Through the processing of more than 100GB of public data via Apache Spark and R Language, as well as using the Random Machines method, the results show important explanatory variables related to the prediction of the municipal usage rate of Bolsa Família in the state of Bahia.
Keywords: *Bolsa Família. Spark. R. Big Data.*

Um estudo preditivo via Máquinas Aleatórias da taxa de utilização municipal do Programa Bolsa Família no estado da Bahia

MATEUS MAIA

Mestre em Matemática e graduado em Geofísica, pela Universidade Federal da Bahia (UFBA). mateusmaia11@gmail.com

MARIANA YUKARI NOGUTI

Especialista em Ciência de Dados e mestranda em Informática, pela Universidade Federal do Paraná (UFPR). Estatística do Ministério Público do Paraná (MPPR). mariana.yn7@gmail.com

ANDERSON ARA

Doutor e mestre em Estatística, pela Universidade Federal de São Carlos (UFSCar). Professor do Departamento de Estatística da Universidade Federal da Bahia (UFBA). anderson.ara@ufba.br

O PROGRAMA Bolsa Família (PBF) é um programa de transferência de renda nacional cujo objetivo é o combate à fome, à pobreza e à desigualdade social através de um repasse financeiro associado a condicionalidades que garantam o acesso aos equipamentos e serviços públicos nas áreas da saúde, educação, assistência social e segurança alimentar. Além disso, procura estabelecer meios para que as famílias em situação de vulnerabilidade sejam capazes de restabelecer sua inclusão social.

As famílias elegíveis são aquelas em situação de pobreza ou extrema pobreza, de acordo com faixas de renda mensal *per capita* estabelecidas pelo governo federal. Além do proposto, as famílias cadastradas devem atender a algumas condicionalidades para o recebimento do benefício, como frequência escolar verificada para crianças e adolescentes entre 6 e 17 anos, calendário vacinal em dia para crianças de 0 a 6 anos e acompanhamento pré-natal junto aos

São encontrados diversos estudos que refletem especificamente a importância do PBF no estado da Bahia como forma de impactar diretamente os indicadores sociais em nível nacional

estabelecimentos de saúde para gestantes. Desse modo, enquanto a transferência monetária atua diretamente no acréscimo de renda familiar, as condicionalidades buscam reduzir a situação de vulnerabilidade familiar através de um acompanhamento adequado nas áreas da saúde e educação. Atualmente, mais de 13,9 milhões de famílias são atendidas pelo Bolsa Família em todo o País.

Desde sua criação, em 2003, diversos estudos foram realizados no intuito de identificar os impactos do Programa Bolsa Família no Brasil. Um artigo do Banco Mundial analisou 47 estudos publicados entre 2009 e 2017 e identificou impactos positivos do PBF na área da saúde, tais como o aumento na utilização de serviços de saúde, aumento na qualidade nutricional e segurança alimentar das famílias beneficiárias e diminuição da taxa de mortalidade infantil (VIANA *et al*, 2018). Em outro artigo sobre a contribuição do PBF para a redução da pobreza no Brasil, os autores apontam para a redução do percentual populacional em situação de pobreza e extrema pobreza devido à expansão do programa ao longo dos anos, conjuntamente com a ampliação na oferta de serviços públicos e abastecimento de água, especialmente em pequenos municípios do Norte e Nordeste. Outro impacto positivo é identificado em município pequenos, cuja economia gira predominantemente na oferta de serviços. Nessas localidades, a renda do Bolsa Família tem grande impacto na dinâmica do comércio local, conforme aponta um artigo do jornal O Estado de S. Paulo (PAAP; PEREIRA, 2017).

Cumpre destacar a importância do programa especialmente para as regiões Norte e Nordeste do país, em que a maior parte das famílias beneficiárias reside em municípios com IDH-M inferior à média nacional, indicando uma situação de vulnerabilidade geograficamente generalizada, enquanto no restante do país encontram-se municípios com bons indicadores, porém com presença de bolsões de pobreza em seu território (MARQUES, 2006). Mais especificamente no estado da Bahia, verifica-se um impacto altamente significativo do programa em razão de ser este o maior estado do Nordeste, com alto quantitativo de famílias em situação de pobreza e extrema pobreza e maior taxa de utilização do benefício em todo o país. Por esses motivos, são encontrados diversos estudos que refletem especificamente a importância do PBF no estado da Bahia como forma de impactar diretamente os indicadores sociais em nível nacional. Pode-se citar um estudo realizado pelo IPEA (SOUZA, 2013) sobre o impacto do programa de transferência de renda na diminuição da desigualdade e dos percentuais de pobreza na Bahia, tendo sido observada uma redução em geral pela metade no período analisado de 2003 a 2011.

Depreende-se das análises acima que o PBF impacta positivamente diversos aspectos da sociedade, desde a melhoria direta nos indicadores

das áreas da educação e saúde até a melhoria econômica das famílias beneficiadas e dos municípios, de modo que a correlação entre o programa e indicadores sociais diversos é claramente estabelecida na literatura. Portanto, é imprescindível aos gestores públicos uma análise detalhada dos dados do programa juntamente com informações e indicadores sociais, a fim de compreender a melhor forma de gerenciar o programa em benefício da população. Como aponta Jannuzzi (2006) em seu livro "Indicadores Sociais no Brasil", é essencial o uso de indicadores no planejamento e implementação de políticas públicas, sendo que a existência de um sistema de indicadores sociais adequado contribui para o sucesso do programa implementado, a partir do diagnóstico e monitoramento de ações e resultados de forma técnica e objetiva.

É imprescindível aos gestores públicos uma análise detalhada dos dados do programa juntamente com informações e indicadores sociais, a fim de compreender a melhor forma de gerenciar o programa em benefício da população

Considerando esses apontamentos, pretende-se neste estudo efetuar uma análise dos municípios do estado da Bahia quanto à taxa de utilização do Programa Bolsa Família, associando-a diretamente aos indicadores sociais das localidades analisadas. O objetivo aqui é pontuar a associação existente entre essas duas fontes de informação, possibilitando reflexões sobre a gestão do recurso público analisado em favor da sociedade. Como exposto anteriormente, o PBF é direcionado a famílias em situação de pobreza e extrema pobreza, de modo que a taxa também pode ser interpretada como um indicativo de vulnerabilidade social do local em questão.

Cumpre destacar também o papel primordial da Lei de Acesso à Informação para a viabilização deste e similares estudos, sem a qual não seria possível a coleta de informações públicas para a análise dos fenômenos pontuados. O planejamento e gestão públicos dependem de fontes objetivas, análises de fatos e estudos técnicos bem fundamentados, obtidos através de dados que demonstrem a realidade dos diferentes municípios do Brasil, bem como possibilitem avaliar o impacto direto que as ações e programas têm sobre eles. Atualmente, a administração pública em geral não possui um repositório consistente e completo de dados abertos de fácil acesso, dificultando a tomada de decisão voltada a dados, sendo compartilhadas apenas algumas informações referentes às políticas públicas monitoradas.

Considerando o contexto da utilização de dados públicos no direcionamento da administração públicas, faz-se necessária a utilização de ferramentas que permitam a exploração e o entendimento de um volume de dados da ordem de grandeza de milhões de indivíduos. Nesse sentido, o uso de conceitos e fundamentos de Big Data tornaram-se pré-requisitos no desenvolvimento de uma administração pública eficaz e baseada em dados (MCNEELY; HAMN, 2014). O aprendizado estatístico de máquina, também se faz presente no desenvolvimento das políticas públicas, uma vez que, através dele, é possível automatizar o processo de tomada de

É necessária a aplicação de ferramentas próprias que possam gerar escalabilidade horizontal do sistema computacional, permitindo, uma vez agregadas, o processamento de um grande volume de dados em sistemas de computação paralela

decisão (ACKERMANN *et al.*, 2018), além de possibilitar o fornecimento de *insight* e obtenção de informações a partir de grandes volumes de dados, auxiliando o processo de tomada de decisão por parte dos órgãos governamentais.

Especificamente no campo de aprendizado de máquina – *machine learning* – os métodos de combinação, os quais são métodos que utilizam mais de um modelo base na composição de um modelo final para prever novas observações, têm ganhado força devido à alta capacidade preditiva, quando comparada com abordagens tradicionais que utilizam um único classificador. Dessa forma, no presente estudo optou-se por utilizar o método de Máquinas Aleatórias de Regressão, pertencente à classe de algoritmos que utilizam a combinação dos métodos de máquina de vetores de suporte (CORTES; VAPNIK, 1995) e possui uma alta capacidade de generalização (ARA *et al.*, 2020; MAIA, 2020), de modo a produzir melhores estimativas quando comparado aos métodos tradicionais.

O artigo está organizado conforme escopo a seguir: na primeira seção é descrito o ambiente computacional que possibilita a manipulação e a exploração do grande volume de dados – Big Data – do Programa Bolsa Família no estado da Bahia. Na seção seguinte, é descrito e apresentado o modelo de aprendizado estatístico que foi utilizado para predição e seleção de variáveis da taxa de utilização municipal do Programa Bolsa Família. Na seção de Resultados e Discussões, são apresentadas as informações obtidas através das modelagens e mineração das informações. Por fim, a última seção apresenta as considerações finais deste artigo.

AMBIENTE COMPUTACIONAL

Big Data é o termo que descreve um grande volume de dados – estruturados e não estruturados – que pode ser analisado para obter informações que levam a melhores decisões e movimentos estratégicos em diversas áreas (SAS, 2020). Ainda que o termo "grande" possa ser subjetivo e difícil de definir, o termo Big Data pode ser considerado quando o volume de dados excede a capacidade de processamento convencional dos sistemas de bancos de dados (DUMBILL, 2012).

Neste sentido, é necessária a aplicação de ferramentas próprias que possam gerar escalabilidade horizontal do sistema computacional, permitindo, uma vez agregadas, o processamento de um grande volume de dados em sistemas de computação paralela. Este procedimento pode ser realizado em uma máquina local ou através da construção de *clusters* computacionais, os quais consistem em computadores interconectados que trabalham em conjunto, de modo que, em muitos aspectos, podem ser considerados como um único sistema.

Dentre as principais ferramentas utilizadas no processamento de Big Data, pode-se citar o processamento MapReduce, o Hadoop e a estrutura HDFS. MapReduce é um modelo de programação proposta pela Google (DEAN; GHEMAWAT, 2008) para processar grandes conjuntos de dados. Os usuários especificam uma função de mapeamento que gera chaves intermediárias e uma função de redução, a qual mescla todos os valores intermediários associados à mesma chave. O Apache Hadoop é um projeto da Apache Software Foundation lançado em 2008, sendo uma plataforma de software em Java de computação distribuída voltada para *clusters* e processamento de grandes volumes de dados, com atenção a tolerância a falhas. O Hadoop Distributed File System (HDFS) foi proposto pela equipe do Yahoo (SHVACHKO *et al*. 2010) para armazenar e transmitir conjuntos de dados muito grandes de maneira confiável.

Mesmo que as aludidas ferramentas sejam desenvolvidas para uso em clusters computacionais, também é possível utilizá-las em uma máquina local

Apache Spark

O Apache Spark é um *framework* para computação em *cluster* para processamento de dados em larga escala. Também foi um projeto da Apache Software Foundation, porém escrito em linguagem Scala e com modelo de execução DAG (*Directed Acyclic Graph*), o qual oferece flexibilidade e desempenho satisfatórios, sendo mais maleável e permitindo recursos diversos voltados para o processamento de dados (CHAMBERS; ZAHARIA, 2018). O Spark tem um modelo de programação semelhante ao MapReduce, mas o estende com uma abstração de compartilhamento de dados chamada "Conjunto de dados distribuído e resiliente" (Resilient Distributed Datasets – RDDs) (ZAHARIA *et al*., 2016). Além disso, o Spark expõe os RDDs por meio de uma API de programação funcional em Scala, Java, Python e R, nas quais os usuários podem simplesmente transmitir funções locais para execução em cluster.

Linguagem R e Spark

O pacote sparklyr (LURASCHI, *et al*. 2018) permite a integração da Linguagem R com o Apache Spark, possibilitando a escrita de um código mais direto para o processamento de grandes bases de dados. De acordo com Shah e outros (2017), o Sparklyr oferece mais funcionalidades e é consideravelmente mais rápido comparado ao seu antecessor SparkR, interface desenvolvida pela Databricks, empresa fundada pelos criadores originais do Apache Spark (DWOSKIN, 2016). O pacote permite transformações de dados baseadas na sintaxe *dplyr*, em conjuntos de dados em ambiente Spark, podendo, ainda, ser utilizado na implantação de tarefas locais ou remotas.

Mesmo que as aludidas ferramentas sejam desenvolvidas para uso em clusters computacionais, também é possível utilizá-las em uma máqui-

Os modelos de vetores de suporte têm características particulares responsáveis pelo seu ótimo desempenho

na local, sendo extremamente úteis para o processamento de grande volume de dados em um computador pessoal, sem a necessidade da utilização de *clusters* ou processamento em nuvem. Nesse sentido, utilizando o Sparklyr e uma única máquina pessoal, este artigo considera o processamento inicial de cerca de 1,2 bilhão de observações relativas ao Programa Bolsa Família, totalizando um tamanho de arquivo com mais de 100 GB.

MÁQUINAS ALEATÓRIAS

Máquinas Aleatórias é um algoritmo de aprendizado estatístico supervisionado, o qual utiliza uma abordagem de combinação de modelos de vetores de suporte (ARA *et al.*, 2019; MAIA, 2020), e abrange o contexto de tarefas de classificação e regressão. Os modelos de máquina de vetores de suporte (MVS) foram inicialmente propostos por Vapnik (CORTES; VAPNIK, 1995) no contexto de classificação e posteriormente foram adaptados para casos de regressão (DRUCKER *et al.*, 1997). Os modelos de vetores de suporte têm características particulares responsáveis pelo seu ótimo desempenho. Um deles se baseia no princípio de Minimização de Risco Estrutural (RME) para minimizar o erro de generalização. Portanto, teoricamente, garante-se que se atinge o mínimo global, enquanto outros algoritmos como Redes Neurais Artificiais (RNA) podem ser capturados em mínimos locais. Outra característica importante do modelo SVM é a base da teoria do aprendizado estatístico de máquina, garantindo a aprendizagem e a generalização do modelo (CORTES; VAPNIK, 1995). Utilizando estas propriedades, as Máquinas Aleatórias de Regressão utilizam uma nova combinação destes modelos, baseando-se no princípio de *bagging* proposto por Breiman (BREIMAN, 1996), através de uma seleção aleatória ponderada das funções de kernel (Quadro 1) que são utilizadas em cada um dos regressores base.

Quadro 1
Funções de kernel para regressões bases de máquina de vetores de suporte

Kernel	Função	Parâmetros
Linear	$\gamma(x \cdot y)$	$\gamma > 0$
Polinomial	γ	$\gamma > 0, d \in \{2,3,\ldots\}$
Gaussiano	$e^{-\gamma \vee x-y \vee^{2}}$	$\gamma > 0$
Laplaciano	$e^{-\gamma \vee x-y \vee}$	$\gamma > 0$

Fonte: Elaborado pelos autores.

Os métodos de combinação – também conhecidos com métodos *ensemble* – destacam-se principalmente devido à sua alta capacidade preditiva (SAGI; ROKACH, 2018), ou seja, sua alta capacidade de generalização resulta em predições mais assertivas para novas observações. As

Florestas Aleatórias propostas por Breiman (BREIMAN, 2001), um dos algoritmos atuais mais populares e robustos devido à sua flexibilidade e acurácia, também pertencem a esta classe de modelos *ensemble*. Nesse sentido, Máquinas Aleatórias de Regressão apresenta-se como uma nova e recente aplicação da combinação de modelos, tendo demonstrado boa performance tanto em experimentos de simulação quanto em uma extensa aplicação em diversos *benchmarkings* de dados reais (ARA *et al.* 2020), mostrando-se superior ao MVS e à sua respectiva abordagem de *bagging*, que tradicionalmente é um método de aprendizado de máquina com alta capacidade preditiva.

Máquinas Aleatórias de Regressão apresenta-se como uma nova e recente aplicação da combinação de modelos, tendo demonstrado boa performance tanto em experimentos de simulação quanto em uma extensa aplicação em diversos benchmarkings de dados reais

Dado um conjunto de treinamento $\{(xi, yi)\}$ com $xi \in R^p$ e $yi \in R$, $\forall i=1,...,n$; o método de Máquinas Aleatórias de Regressão é inicializado através do treinamento de modelos únicos $hr(x)$ (x), em que $r = 1,..., R$; sendo R o número total de funções diferentes do kernel que podem ser usadas em modelos de regressão de vetores de suporte (Quadro 1). Por exemplo, se $R = 4$, uma opção possível é definir h1 como MVS com kernel Linear, h2 como MVS com kernel Polinomial, h3 como MVS com kernel Gaussiano e h4 como MVS com kernel Laplaciano.

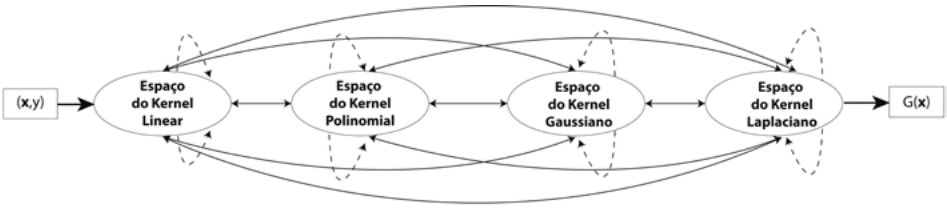
Cada modelo é validado para o conjunto de testes $\{(xt, yt)\}$ e a raiz do erro quadrático médio (REQMr), que iremos chamar de δi, é calculado para cada modelo $hr(x)$, em que R é o número de funções do kernel que serão usadas. O REQMr é uma métrica de avaliação tradicionalmente usada para verificar a qualidade da predição, uma vez que calcula a raiz do valor médio do quadrado das diferenças entre os valores preditos e os observados. Como o intervalo da variável dependente na regressão (y) é amplo, o vetor da raiz significa quadrados δi é dividido pelo seu desvio para padronizar o erro. Posteriormente, as probabilidades da amostra, λi, são calculadas pela Equação (1) para cada função do kernel,

$$\lambda_i = \frac{e^{-\beta\delta_i}}{\sum_{j=1}^{B} e^{-\beta\delta_j}},\ i=1,\ \dots\ ,\ B \qquad (1)$$

onde β é um coeficiente de ponderação do erro. Em seguida são geradas B amostras *bootstrap*, e então é estimado um modelo de máquina de vetores de suporte *g*i para cada uma dessas amostras, onde em cada uma será selecionada aleatoriamente uma função de kernel a ser utilizada no MVS. Essa seleção aleatória permite um salto em diferentes espaços dimensionais entre cada um dos modelos *bootstrap*, explorando assim o espaço de representações das observações, obtendo diferentes características e melhorando a capacidade de generalização do modelo final. A Figura 1 representa esquematicamente essa transição.

O modelo de MVS pode ser interpretado como um caso particular do Máquinas Aleatórias, onde há apenas uma repetição e a probabilidade de selecionar uma determinada função de kernel é igual a 1

Figura 1
Representação dos saltos entre diferentes espaços dimensionais através das funções de kernel



Fonte: Elaborado pelos autores.

Algoritmo 1. Máquinas Aleatórias de Regressão

Entrada: Conjunto de dados de treinamento e teste para regressão, sendo Y a variável resposta e X o vetor de p variáveis explicativas; hi funções de kernel que serão utilizadas; B número de funções de amostras *bootstrap*;

1. Para cada função de kernel hi
   1.1. Calcula o modelo hi
   1.2. Calcula a probabilidade λi
2. Gera B amostras *bootstrap*.
3. Para cada amostra *bootstrap*
   3.1. Calcula um modelo de máquina de vetores de suporte gb utilizando uma seleção aleatória da função de kernel com base na probabilidade λ.
   3.2. Associa um peso wb para este modelo utilizando a raiz do erro quadrático médio em uma amostra *Out of the Bag (OOBG)*
4. Calcula o valor predito final através a combinação de todas estimativas através da função G.

Saída: Valores preditos para o conjunto teste e treinamento

Posteriormente, o peso da estimativa de cada um dos modelos é dado pela Equação 2, onde representa a raiz do erro quadrático médio de cada um dos modelos de cada amostra *bootstrap* sobre as amostras *Out of the Bag (OOBG)* (BREIMAN, 2001),

$$w_i = \frac{e^{-\beta \Delta_i}}{\sum_{j=1}^{B} e^{-\beta \Delta_j}} \; i=1, \ldots, B \qquad (2)$$

o fim do treinamento de todos os modelos a estimativa é combinada através da média ponderada dos valores representado pela Equação (3).

$$G(x_i) = \sum_{j=1}^{B} w_j g_j(x_i) \; i=1, \ldots, n \qquad (3)$$

O método é Máquinas Aleatórias de Regressão é sumarizado no Algoritmo 1. É interessante observar que o modelo de MVS pode ser interpretado como um caso particular do Máquinas Aleatórias, onde há apenas uma repetição e a probabilidade de selecionar uma determinada função de kernel é igual a 1.

Método de seleção de variáveis

Um dos principais problemas na construção de modelos é a escolha entre um grande conjunto de variáveis explicativas no sentido de encontrar um subconjunto de variáveis significativas que devam ser incluídas no modelo final

Um dos principais problemas na construção de modelos é a escolha entre um grande conjunto de variáveis explicativas no sentido de encontrar um subconjunto de variáveis significativas que devam ser incluídas no modelo final. Como consequência, muitos autores geralmente realizam seleção de variáveis apenas uma vez, com aplicação de todos os dados disponíveis, podendo induzir a uma subestimação drástica do erro de predição e, assim, levar a relatar enganosamente o poder preditivo. Além disso, existem diferentes algoritmos de seleção de variáveis que variam de acordo com critérios de testes ou medidas de precisão, desempenho ou, ainda, no método de validação. Na regressão linear múltipla, por exemplo, o método *stepwise* é o mais amplamente difundido método seleção, o qual adiciona ou remove variáveis explicativas, geralmente por meio de uma série de testes F ou testes T, tendo como variantes os métodos de seleção *forward* e *backward*, caracterizados por adicionar e remover variáveis uma a uma, respectivamente (MONTGOMERY; RUNGER, 2009). Métodos menos comuns incluem todas as regressões possíveis, como o método Cp de Mallow's (MALLOWS, 1973). Na classificação binária, por exemplo, é usual selecionar variáveis explicativas de acordo com o p-valor obtido em testes para duas amostras independentes, tais como teste T ou testes não paramétricos (BOULESTEIX, 2007; DETTLING; BUËHLMANN, 2003). Neste artigo, foi utilizada uma variação do método de seleção de variáveis WilcoxCV proposto por Boulesteix (2007) e já utilizado na seleção de variáveis em métodos de máquina de vetores de suporte (BALLABIO; STERLACCHINI, 2012; FRONZA *et al*. 2013) . Este método se baseia em classificação binária, na utilização do teste de soma de postos de Wilcox, também conhecido como teste de Mann-Whitney ou teste de Wilcox para duas amostras independentes, e em simulações de Monte Carlo.

O método utilizado neste artigo é uma adaptação do método *forward* aplicado à regressão via Máquinas Aleatórias, bem como na utilização do teste de postos sinalizados de Wilcoxon, também conhecido como teste de Wilcox para duas amostras pareadas, e em simulações de Monte Carlo via validação através do método *hold-out* repetido. A estrutura *forward* baseia-se em incluir ordenadamente variáveis uma a uma, com o intuito de aumentar a capacidade preditiva da regressão, mensurada via REQMr. A verificação de mudança significativa da capacidade preditiva é verificada, a um nível alfa de significância previamente estabelecido, via teste de postos sinalizados de Wilcoxon, o qual é adequadamente aplicado a amostras pareadas durante o processo iterativo de seleção. Além disso, o método de validação *data splitting* via *hold-out* é um dos mais amplamente utilizados, sendo sua versão repetida com melhores propriedades de estimação do erro de predição (TANTITHAMTHAVORN

et al., 2016). O método *forward* Wilcox pareado via *hold-out* repetido é apresentado no Algoritmo 2.

Algoritmo 2. Forward Wilcox pareado via hold-out repetido
Entrada: Conjunto de dados para regressão, sendo Y a variável resposta e X o vetor de p variáveis explicativas; REP o número de repetições; %treinamento proporção dos dados utilizados como treinamento; alfa = nível de significância adotado
1. Para cada repetição 1,.., REP 1.1. Dividir o conjunto em treinamento e teste via %treinamento 1.2. Para cada variável explicativa i, i=1,...,p 1.2.1. Realizar o ajuste do modelo de regressão de Y sobre Xi via algoritmo X e utilizando os dados de treinamento 1.2.2. Calcular o REQMr para dados de teste 2. Ordenar as variáveis via REQMr 3. Iniciar o modelo com o menor REQMr 4. Enquanto p.valor < alfa 4.1. Incluir no modelo a nova variável explicativa pela ordenação de menores REQMr 4.2. Realizar o ajuste para as repetições 1,.., REP 4.3. Testar a hipótese da inclusão da nova variável via teste de Wilcoxon pareado e REQMr nos dados de teste 4.4. Caso a não rejeição da hipótese via p.valor, manter a nova variável no modelo
Saída: Variáveis selecionadas

RESULTADOS E DISCUSSÃO

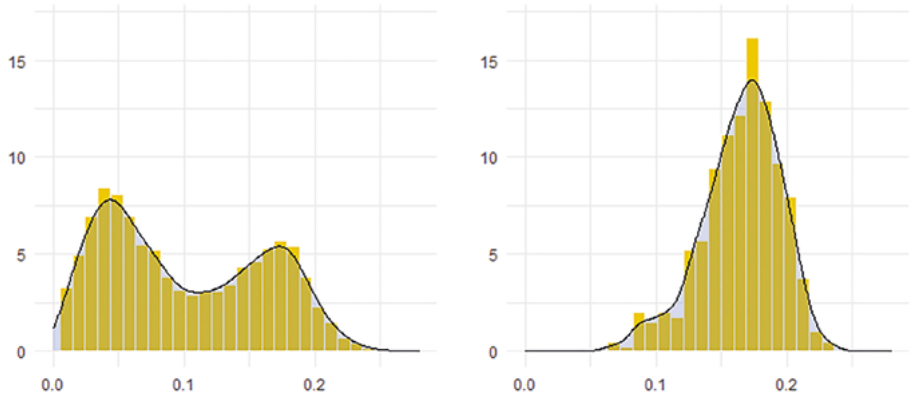
Descrição e origem dos dados

Os dados utilizados são referentes aos pagamentos mensais do Programa Bolsa Família no período de janeiro de 2013 a dezembro de 2019 e foram extraídos do Portal Brasileiro de Transparência (PORTAL DA TRANSPARÊNCIA, 2020). Os arquivos originais estão disponibilizados no formato CSV e se referem aos pagamentos nominais do benefício por mês, discriminados por município e valor recebido. No total e para o país inteiro, existem aproximadamente 15 milhões de pagamentos por mês, os quais, agregados, totalizam um arquivo de tamanho 111,6 GB, com 1,26 bilhão de observações. Neste sentido, a taxa de utilização do bolsa família por município é obtida a partir da definição abaixo

$$TaxaBF = número de beneficiários / total populacional$$

sendo o total populacional obtido como o número de habitantes do município no último Censo (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2020). Esta taxa foi sumarizada, por município, via média do horizonte dos 84 meses em estudo. Assim, no sentido de verificar a explicação da taxa municipal a partir de informações municipais, são consideradas as variáveis semelhantes ao estudo realizado por Costa e outros (2018), porém em um contexto municipal. As variáveis explicativas foram coletadas através do Sistema Nacional de Informações de Gênero (SNIG), o qual integra o projeto de estruturação de um amplo Programa de Estatísticas de Gênero no Instituto Brasileiro de Geografia e Estatística (2020). As variáveis utilizadas estão detalhadas no Quadro 2.

Figura 2
Histograma da taxa de utilização municipal do Programa Bolsa Família



No Brasil | No Estado da Bahia

Fonte: Elaborado pelos autores.

Quadro 2
Descrição das variáveis utilizadas no presente artigo

Variável	Descrição
TAXA_BF	Taxa de utilização municipal média do Programa Bolsa Família entre 2013 e 2019
T_MULHERES	Razão entre total do sexo feminino pelo total de habitantes
T_BRANCOS	% de pessoas de cor ou raça branca
T_UNIRESPH	% de unidades domésticas com único responsável homem com cônjuge e filhos e/ou outros parentes
T_ANALF	Taxa de analfabetismo da população de 18 anos ou mais de idade
E_ANOSESTUDO	Expectativa de anos de estudo
P_FORMAL	Grau de formalização dos ocupados – 18 anos ou mais
T_DES18M	Taxa de desocupação – 18 anos ou mais
T_PPFAMILIA	Razão entre o número de habitantes e o número de famílias
T_MULHERCHEF	Proporção de famílias com mulheres responsáveis pela família (%)
T_PEA	Razão entre população economicamente ativa de 18 anos ou mais pela população
T_DEF_AUD	Proporção de pessoas com deficiência auditiva
T_DEF_MOT	Proporção de pessoas com deficiência motora
T_DEF_VIS	Proporção de pessoas com deficiência visual
T_DEF_MENT	Proporção de pessoas com deficiência mental/intelectual
T_AGUA	% da população em domicílios com água encanada
T_BANAGUA	% da população em domicílios com banheiro e água encanada
T_DENS	% da população em domicílios com densidade > 2
T_LIXO	% da população em domicílios com coleta de lixo
T_LUZ	% da população em domicílios com energia elétrica
AGUA_ESGOTO	% de pessoas em domicílios com abastecimento de água e esgoto inadequados
PAREDE	% de pessoas em domicílios com paredes inadequadas
EMP	% de empregadores – 18 anos ou mais
GINI	Índice de Gini

Fonte: SNIG/IBGE – Censo 2010. Elaborado pelos autores.

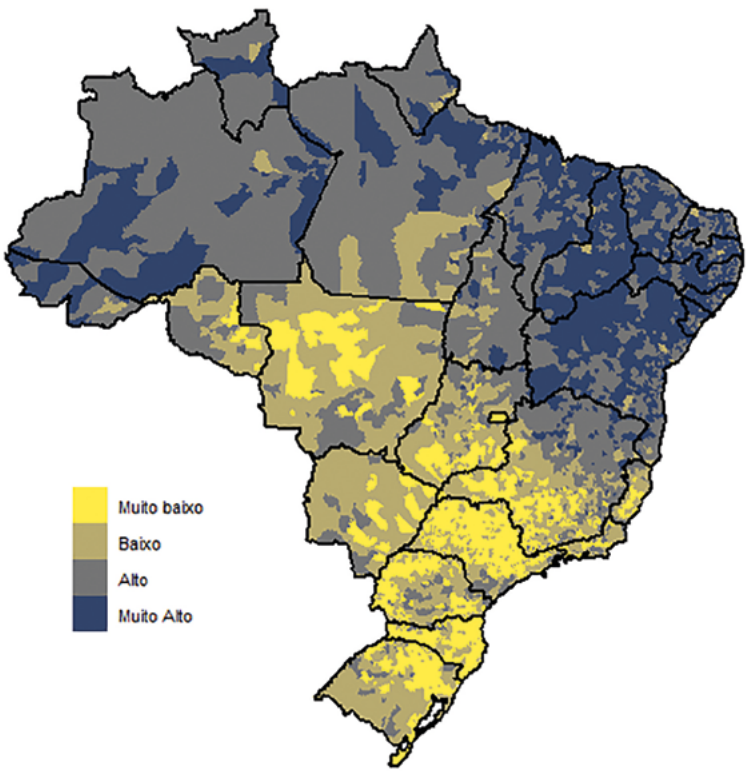
A distribuição da taxa de utilização municipal é exibida na Figura 2, sendo possível observar um comportamento bimodal para o Brasil (a), identificando a existência de pelo menos dois tipos de municípios, com alta e baixa taxa de utilização do PBF. Da mesma forma, a Figura 2 (b)

As informações obtidas de maneira mais nítida através da visualização de um grande volume de dados permitem um direcionamento mais eficaz das políticas públicas, considerando as regiões mais vulneráveis como alvo de programas socioeconômicos, visando à redução das desigualdades

exibe a distribuição para os municípios baianos, os quais apresentam o segundo comportamento nacional de alta utilização do PBF.

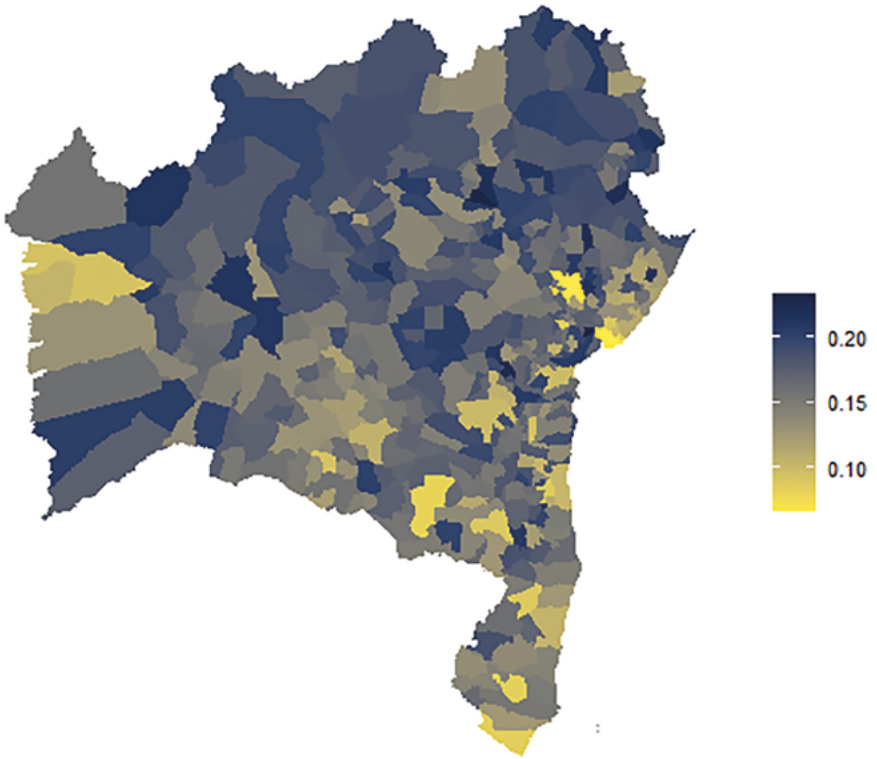
A partir da Figura 3, a qual corrobora com a Figura 2, observa-se uma evidente desigualdade entre as regiões do Brasil, com a categorização da taxa de utilização municipal do PBF em quatro categorias a partir dos quartis, sendo elas respectivamente: muito baixo, baixo, alto e muito alto. As regiões Nordeste e Norte concentram um grande número de municípios que possuem uma taxa de utilização alta e muito alta, enquanto as regiões Centro-Oeste, Sudeste e Sul apresentam um maior número de municípios classificados como baixo e muito baixo. As informações obtidas de maneira mais nítida através da visualização de um grande volume de dados permitem um direcionamento mais eficaz das políticas públicas, considerando as regiões mais vulneráveis como alvo de programas socioeconômicos, visando à redução das desigualdades. Dentro do estado da Bahia, ainda é possível também observar esse fenômeno, apesar da predominância das categorias alto e muito alto, visto que a maior parte da região Sul do estado apresenta uma taxa de utilização municipal menor quando comparado ao Centro-Oeste, Norte e Nordeste.

Figura 3
Categorias da taxa de utilização do Programa Bolsa Família por município



Fonte: Elaborado pelos autores.

Figura 4
Taxa média de utilização do Programa Bolsa Família, por município – Bahia – 2013-2019



Fonte: Elaborado pelos autores.

Inicialmente, o modelo foi aplicado de maneira univariada para cada uma das variáveis explicativas

Modelagem via Máquinas Aleatórias

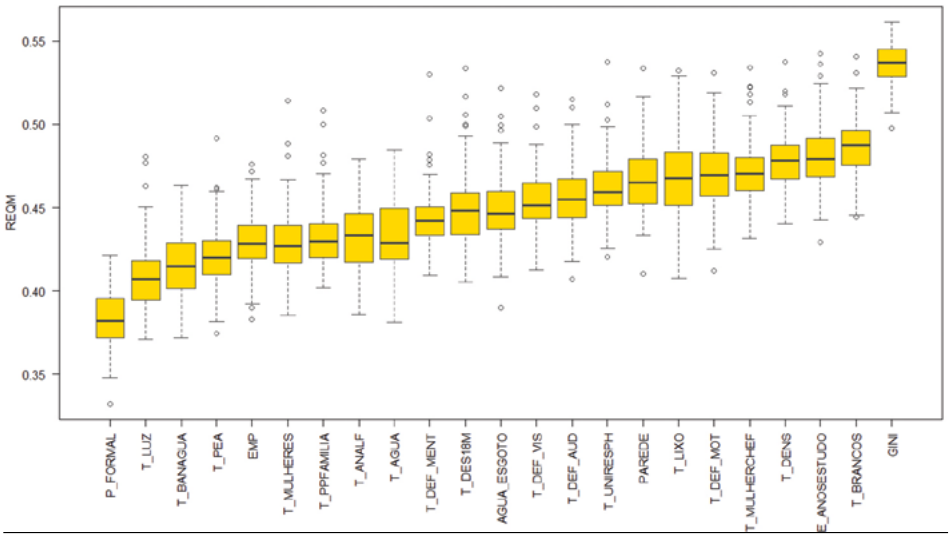
Considerando a base de dados do Programa Bolsa Família por município brasileiro, aplicou-se o modelo de Máquinas Aleatórias de Regressão para estimar a taxa de utilização do Bolsa Família nos municípios baianos. As covariáveis analisadas são apresentadas no Quadro 2, com os parâmetros do modelo: 100 amostras *bootstrap,* 4 funções de kernel (Quadro 1) com respectivos hiper-parâmetros d=2, , e coeficiente de correlação . Para a validação do modelo, foi utilizada a técnica de *hold-out* repetido, com 100 repetições e a proporção entre base de treinamento-teste de 80-20% respectivamente. A métrica de avaliação preditiva foi a raiz do erro quadrático médio (REQM).

Inicialmente, o modelo foi aplicado de maneira univariada para cada uma das variáveis explicativas. O resultado obtido pode ser visualizado através da Figura 5, em que se apresentam os boxplots do REQM para cada uma das variáveis explicativas de maneira ordenada, do menor para o maior. É possível perceber que variáveis como emprego formal,

Uma vez que estas variáveis foram obtidas podem ser utilizadas no desenvolvimento de políticas públicas que busquem uma melhor aplicação dos recursos públicos e, especificamente, na gestão do Programa Bolsa Família nos municípios

domicílios com energia elétrica e saneamento básico apresentam um menor valor de erro associado à predição, demonstrando que essas covariáveis, associadas ao nível de pobreza de um município, possuem um poder explicativo maior em relação à taxa de utilização municipal do Programa Bolsa Família, quando comparado a outras características como o índice GINI e anos de estudo.

Figura 5
Boxplot da Raiz do Erro Quadrático dos modelos de Máquinas Aleatórias de Regressão univariados



Fonte: Elaborado pelos autores.

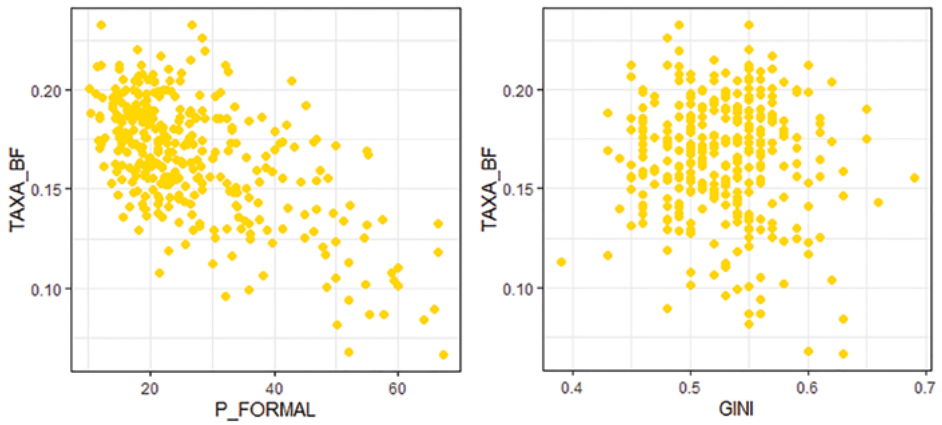
A relação entre o Índice de Gini e a taxa de utilização do Programa Bolsa Família, verificada via teste de coeficiente de correlação de Spearman (r), não é significativa a 5% de significância (p-valor = 0.8556).

Utilizando o ranking de variáveis da Figura 6, foi aplicado um processo *forward* via Wilcox pareado, através de *hold-out* repetido na geração dos modelos de Máquinas Aleatórias de Regressão, a 5% de significância, adicionando-se sequencialmente as variáveis explicativas até que não houvesse acréscimo significativo do REQM com o incremento da nova variável. A partir desse processo, foram selecionadas as variáveis P_FORMAL, T_LUZ, T_BANAGUA, T_PEA e EMP, as quais, a partir da segunda variável, com p-valores respectivos de <,001; 0,0369; 0,0144 e 0,0477 para o método de seleção de Wilcox pareado. A variável T_MULHERES foi descartada com p-valor igual a 0,6859.

Uma vez que estas variáveis foram obtidas podem ser utilizadas no desenvolvimento de políticas públicas que busquem uma melhor aplicação dos recursos públicos e, especificamente, na gestão do Programa Bolsa Família nos municípios. Criar ações que facilitem a geração de empre-

gos, bem como a situação para um ambiente favorável no surgimento de empregadores, pode ser uma meta a ser alcançada pelo poder público. Além disso, indicadores sociais como porcentagem de domicílios com energia elétrica, banheiro e água encanada apontam que o fator social também é predominante na taxa de utilização municipal.

Figura 6
Scatterplot entre a taxa de utilização do PBF por Emprego Formal e Índice de Gini



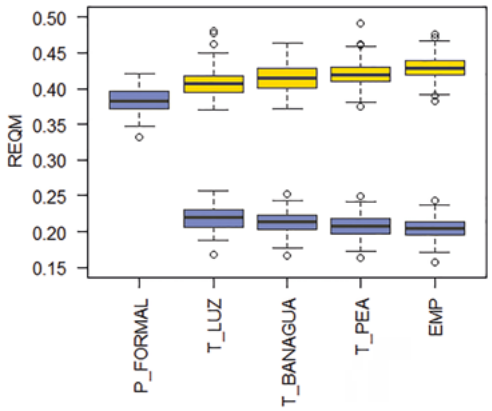
(a)Taxa de utilização x Emprego Formal (a)Taxa de utilização x Índice GINI

Fonte: Elaborado pelos autores.

No sentido de também verificar a relação de dependência das variáveis explicativas com a variável resposta, foram calculados coeficientes de correlação linear de Spearman (r), tendo sido na totalidade significativos a um nível de 5% de significância, bem como apresentaram correlação negativa em todos os casos, indicando uma associação moderadamente ou fracamente linear negativa com a taxa média de utilização do Programa Bolsa Família nos municípios baianos. Desse modo, apresentam-se a seguir os coeficientes em ordem de seleção de variáveis, respectivamente: P_FORMAL (r = – 0,239), T_LUZ (r = – 0,506), T_BANAGUA (r = – 0,283), T_PEA (r = – 0,418) e EMP (r = – 0,257). Destaca-se dos resultados que, além das relações com variáveis demográficas relativas à taxa de população economicamente ativa, há também evidências de que a elevação do grau de formalização dos ocupados e a proporção de empregadores, bem como o aumento no número de domicílios com energia elétrica, banheiro e água encanada, resultaria, de forma moderada ou fraca, em uma menor taxa de utilização municipal do Programa Bolsa Família.

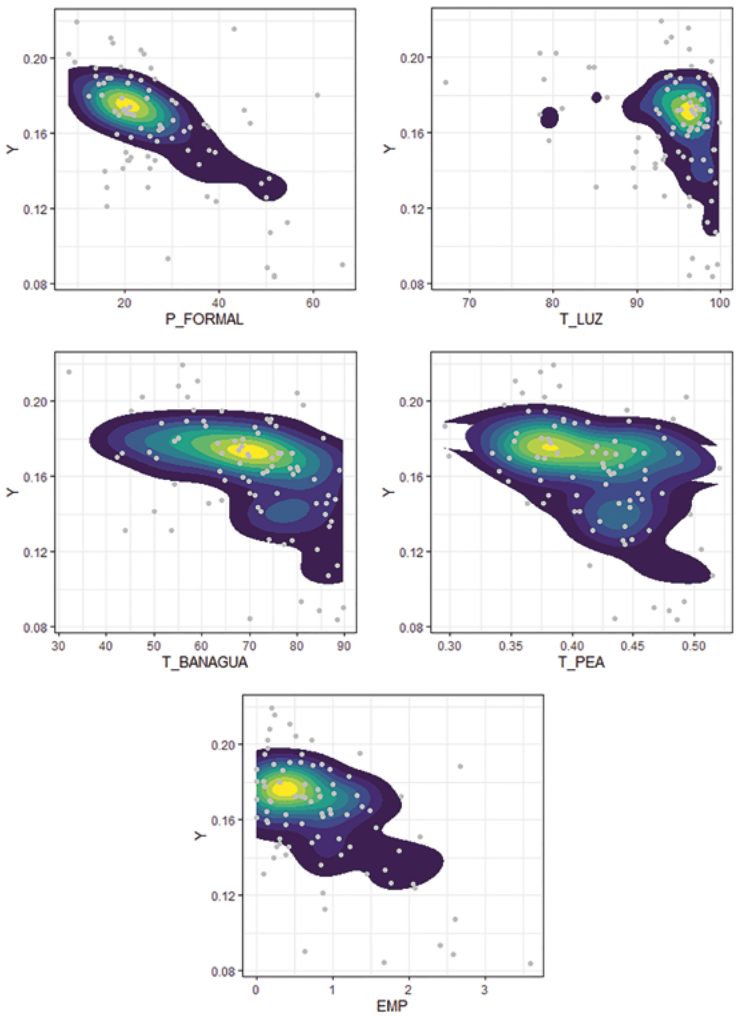
Para visualizar a flexibilidade na predição utilizando o método de Máquinas Aleatórias, a Figura 8 exibe sua superfície de predição final, em curvas de nível, com relação a cada uma das variáveis selecionadas e em comparação com uma amostra teste (observações em cinza). Na imagem, nota-se uma relação negativa entre a variável resposta e as va-

Figura 7
Boxplot da Raiz do Erro Quadrático dos modelos de Máquinas Aleatórias de Regressão de modelos aninhados das variáveis selecionadas em azul, e em univariadas em amarelo



Fonte: Elaborado pelos autores.

Figura 8
Superfície de predição do modelo ajustado pelo método de Máquinas Aleatórias



Fonte: Elaborado pelos autores.

riáveis explicativas que, no entanto, evidencia uma estrutura não linear ou fracamente linear, ainda assim captada de forma geral pelo método de previsão.

Observa-se a evidente desigualdade entre os valores dos indicadores socioeconômicos, destacando-se aqueles referentes à porcentagem de empregos formais, empregadores e à porcentagem dos domicílios com banheiro e água encanada

Tabela 1
Cinco municípios baianos com a maior utilização do Programa Bolsa Família

	Nome	Taxa_bf	P_formal	T_luz	T_banagua	T_pea	Emp
1	Caldeirão Grande	0,233	12,16	92,67	59,97	0,44	0,00
2	Saubara	0,232	26,77	98,18	93,37	0,45	0,45
3	Santa Inês	0,226	28,30	97,59	73,27	0,41	0,44
4	Biritinga	0,220	17,94	95,64	57,70	0,40	0,27
5	Ribeira do Amparo	0,220	9,72	92,96	56,02	0,38	0,20
Média		0,226	18,98	95,41	68,07	0,42	0,27
Desv. P.		0,006	8,38	2,55	15,70	0,03	0,19

Fonte: Elaborado pelos autores.

Tabela 2
Cinco municípios baianos com a menor utilização do Programa Bolsa Família

	Nome	Taxa_bf	P_formal	T_luz	T_banagua	T_pea	Emp
1	Salvador	0,067	67,16	99,85	95,09	0,53	1,87
2	Feira de Santana	0,068	51,94	99,77	86,61	0,51	2,16
3	Vitória da Conquista	0,082	50,21	99,19	87,03	0,49	2,18
4	Teixeira de Freitas	0,084	51,71	99,10	88,33	0,49	3,58
5	Mucuri	0,084	51,55	96,30	70,17	0,43	1,67
Média		0,077	54,51	98,84	85,45	0,49	2,29
Desv. P.		0,009	7,10	1,46	9,20	0,04	0,75

Fonte: Elaborado pelos autores.

A análise descritiva da taxa de utilização municipal do PBF e as variáveis selecionadas nos cinco municípios com a maior e menor taxa estão apresentadas nas Tabelas 1 e 2, respectivamente. Comparando ambas, observa-se a evidente desigualdade entre os valores dos indicadores socioeconômicos, destacando-se aqueles referentes à porcentagem de empregos formais, empregadores e à porcentagem dos domicílios com banheiro e água encanada. Os resultados corroboram com as variáveis selecionadas.

CONSIDERAÇÕES FINAIS

O crescimento exponencial na geração e armazenamento de dados públicos torna necessária utilização de conceitos de Big Data no desenvolvimento de políticas públicas orientadas a dados. Dessa forma, o presente artigo demonstra a utilização de ferramentas recentes como Apache Spark e Linguagem R para manipulação e exploração de grandes volumes de dados públicos, com o intuito de extrair os dados do estado da Bahia a partir de dados nacionais, oriundos de um arquivo

Ao utilizar uma modelagem robusta baseada em métodos de combinação de modelos de máquina de vetores de suporte para regressão através do método de Máquinas Aleatórias, verificou-se que a taxa de utilização municipal na Bahia pode ser explicada através de indicadores demográficos

extraído do Portal da Transparência, de dimensionamento total de 100 GB, além de exibir uma aplicação original de predição da taxa de utilização municipal mensal média dos municípios baianos.

Dentre os principais resultados deste artigo, destaca-se que o estado da Bahia possui uma taxa de utilização municipal do PBF que difere daquela encontrada a nível nacional, em específico se comparada com taxas de outros estados das regiões Sul, Sudeste e Centro-Oeste. Além disso, uma vez que o PBF possui especificações próprias para aquisição do benefício, baseadas em características econômicas e sociais das famílias, pode-se entender a taxa de utilização municipal como um indicador de vulnerabilidade socioeconômica, não representado em outros indicadores sociais, como o Índice de Gini dos municípios baianos. Ainda, ao utilizar uma modelagem robusta baseada em métodos de combinação de modelos de máquina de vetores de suporte para regressão através do método de Máquinas Aleatórias, verificou-se que a taxa de utilização municipal na Bahia pode ser explicada através de indicadores demográficos, como características da população do município (proporção da população economicamente ativa), por indicadores de habitação e urbanismo (proporção de domicílios com energia elétrica e proporção em domicílios com banheiro e água encanada), bem como por indicadores relativos ao trabalho (grau de formalização dos ocupados e proporção de empregadores). Neste sentido, recomenda-se um estudo mais aprofundado sobre o Programa Bolsa Família em municípios baianos com a presença de população infanto-juvenil e de idosos, os quais não estão compreendidos na população economicamente ativa deste estudo. Por fim, em relação às recomendações administrativas voltadas à gestão pública, pontua-se o aumento de investimentos em aspectos urbanísticos e de saneamento básico (água, esgoto e energia elétrica) e de ações que fomentem o aumento da formalização do trabalho e da presença de empregadores formais nos municípios.

Cumpre ressaltar, ainda, que este artigo não leva em consideração o número total de habitantes do município, uma vez que esta variável é utilizada na composição de várias variáveis explicativas aqui utilizadas. Além disso, considera a taxa média da utilização municipal do Programa Bolsa Família no período de janeiro de 2013 a dezembro de 2019, sem considerar a variabilidade temporal da aludida taxa por mês. Neste sentido, outros trabalhos e pesquisas podem ser desenvolvidos para um estudo que considere este aspecto de variabilidade temporal. Ademais, seria viável a utilização da mesma abordagem deste artigo com outros métodos de aprendizado estatístico de máquina, bem como outras variáveis e métodos de seleção de variáveis para predição da taxa de utilização municipal no estado da Bahia, ou ainda para outras regiões do Brasil. Também poderiam ser realizados ajustes nas taxas de utilização

a partir de estimativas populacionais para os anos analisados (2013 a 2019), bem como para as demais variáveis utilizadas.

Por fim, os autores são detentores dos dados agregados por município em todo território nacional e das implementações utilizadas neste artigo, os quais estão disponíveis, mediante a solicitação, para pesquisas futuras e continuação deste trabalho.

REFERÊNCIAS

ACKERMANN, K. *et al* WALSH, J., De Unánue, A., Naveed, H., Navarrete Rivera, A., Lee, S. J., Ghani, R. Deploying machine learning models for public policy: A a framework. In: YIKE, Guo; FAISAL, Faroog. KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: Association for Computing Machinery, 2018. p. 15-22. to *Apache Software Foundation*. Disponível em: <https://www.apache.org/>. Acesso em: 10 jun. 2020.

ARA, A. *et al*. *Random machines*: a bagged-weighted support vector model with free kernel choice. [S. l: s. n], 2019. Disponível em: https://arxiv.org/abs/1911.09411. Acesso em: 18 nov. 2020.

ARA, A MAIA, Mateus; MACÊDO, Samuel; LOUZADA, Francisco. *et al*. Random machines regression approach: an ensemble support vector regression model with free kernel choice. [S. l: s. n], arXiv preprint arXiv:2003.12643, 2020. Disponível em: https://arxiv.org/abs/2003.12643. Acesso em: 18 nov. 2020.

BALLABIO, C.; STERLACCHINI, S. Support vector machines for landslide susceptibility mapping: the Staffora River Basin case study. Italy, *Mathematical. Geosciences*., Italy, v. 44, n. 1, p. 47–70. , Jan. 2012. Disponível em: https://www.researchgate.net/publication/230690362_Support_Vector_Machines_for_Landslide_Susceptibility_Mapping_The_Staffora_River_Basin_Case_Study_Italy. Acesso em: 18 nov. 2020.

BOULESTEIX, A. L.; TUTZ, G. Identification of interaction patterns and classification with applications to microarray data. *Computational Satistics & Data Analysis*, [s. l.], v. 50, n. 3, p. 783-802, Feb. 2006. Disponível em: https://dl.acm.org/doi/10.1016/j.csda.2004.10.004. Acesso em: 18 nov. 2020.

BOULESTEIX, A. Laure. WilcoxCV: an R package for fast variable selection in cross-validation. *Bioinformatics*, [s. l.], v. 23, n. 13, p. 1702-1704, July, 2007. Disponível em: https://academic.oup.com/bioinformatics/article/23/13/1702/223888. Acesso em: 18 nov. 2020.

BREIMAN, L. Bagging predictors. *Machine Learning*, [s. l.], v. 24, n. 2, p. 123-140, Aug. 1996. Disponível em: https://link.springer.com/article/10.1007/BF00058655. Acesso em: 18 nov. 2020

BREIMAN, Leo. Random forests. *Machine Learning*, [s. l.], v. 45, n. 1, p. 5-32, Oct. 2001. Disponível em: https://link.springer.com/article/10.1023/A:1010933404324. Acesso em: 18 nov. 2020.

CENSO DEMOGRÁFICO 2010. Rio de Janeiro: IBGE, 2012.

CHAMBERS, B.; ZAHARIA, M. *Spark*: the definitive guide: Big data processing made simple. [S. l.]: O'Reilly Media, 2018.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, [s. l.] v. 20, n. 3, p. 273-297, 1995. Disponível em: https://link.springer.com/article/10.1007/BF00994018. Acesso em: 18 nov. 2020

COSTA, R. A. *et al*. Impactos do Programa Bolsa Família no mercado de trabalho e na renda dos trabalhadores rurais. *Nova Economia*, Belo Horizonte, v. 28, n. 2, p. 385-416, maio/ago. 2018.

DEAN, J.; GHEMAWAT, S. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, [s. l.], v. 51, n. 1, p. 107-113, Jan. 2008. Disponível em: https://dl.acm.org/doi/10.1145/1327452.1327492. Acesso em: 18 nov. 2020.

DETTLING, M.; BÜHLMANN, P. Boosting for tumor classification with gene expression data. *Bioinformatics*, [S. l.], v. 19, n. 9, p. 1061-1069, Jun. 2003.

DRUCKER, H. *et al*. Support vector regression machines. *Advances in neural information processing systems*, [S. l.], v. 28, n. 7, p. Jan. 779-784, 1997. Disponível em: https://www.researchgate.net/publication/309185766_Support_vector_regression_machines. Acesso em: 28 nov. 2020. p. 155-161.

DUMBILL, E. *Planning for big data*. Sebastopol: O'Reilly, 2012.

DWOSKIN, E. This is where the real action in artificial intelligence takes place. *Washington Post* 9 jun. 2016. Disponível em: http://www.washingtonpost.com/news/. Acesso em: 10 jun. 2020.

FRONZA, I. *et al*.TTI, A., SUCCI, G., TERHO, M., VLASENKO, J. Failure prediction based on log files using random indexing and support vector machines. *Journal of Systems and Software*, [s. l.], v. 86, n. (1, p.), 2-11, 2013.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. - *Sistema Nacional de informações de Gênero*. Disponível em: <https://www.ibge.gov.br/apps/snig/v1/>. Acesso em: 15 maio 2020.

JANNUZZI, P. de M.artino. *Indicadores sociais no Brasil*: conceitos, fontes de dados e aplicações para formulação e avaliação de políticas públicas, elaboração de estudos socioeconômicos. [São Paulo]; Alínea. Editora, 2006. p. 141-141, 2006..

LURASCHI, J. *et al*. *The apache software foundation*: sparklyr: R Interface to Apache Spark. [S. l.: s. n], (2018). (R package version 0.6, 3).

MCNEELY, C. L.; HAHM, J.ong-on. The big (data) bang: Policy, prospects, and challenges. *Review of Policy Research*, [s. l.], v. 31, n. 4, p. 304-310, 2014.

MAIA, M. M. *Máquinas Aleatórias*: o espaço kernel aleatório para construção de um método de combinação de máquinas de vetores de suporte. Dissertação de Mestrado (Matemática com área de concentração em Estatística), - Universidade Federal da Bahia -, Salvador, 2020.

MALLOWS, C. L. Some comments on Cp. *Technometrics*, [s. l.], v. 15, n. 4, p. 661-675, 1973.

MARQUES. R. M. *A importância do Bolsa Família nos municípios brasileiros - segundo estudo*. Brasília: Ministério do Desenvolvimento Social e Combate à Fome, Secretaria de Avaliação e Gestão da Informação, 2006. (Cadernos de estudos desenvolvimento social em debate, 1).

MONTGOMERY, D. C.; RUNGER, G. C. *Estatística aplicada e probabilidade para engenheiros*. 4. ed. Rio de Janeiro: LTC, 2009. 493 p.

PAAP, A. CarolinaC.; PEREIRA, Renée. Bolsa Família evita o colapso de cidades. *O Estado de São Paulo*, São Paulo, n. 45036, 05 fev. /02/2017. Economia, p. B3. Disponível em: <http://www2.senado.leg.br/bdsf/handle/id/529986. Acesso em: 18 nov. 2020.

PORTAL DA TRANSPARÊNCIA (Brasil). *Bolsa Família*: pagamentos. Disponível em: <http://www.portaltransparencia.gov.br/download-de-dados/bolsa-familia-pagamentos/>. Acesso em: 10 maio 2020.

SAGI, O.; ROKACH, Lior. Ensemble learning: *a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, [s. l.], v. 8, n. 4, p. e1249, 2018.

SAS. *Big Data*: what it is and why it matters O que é e qual sua importância?. Disponível em: <https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html>. Acesso em: 10 jun. 2020.

SHAH, P.; HIREMATH, D.; CHAUDHARY, S. Towards development of spark based agricultural information system including geo-spatial data. *In*: IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 2017, [s. l.]. *Anais* [...]. [S. l.]: IEEE, 2017. p. 3476-3481.

SHVACHKO, K. *et al*.The hadoop distributed file system. In: IEEE SsYMPOSIUM ON MASS STORAGE SYSTEMS AND TECHNOLOGIES (MSST), 26., 2010, [s. l.]. *Anais* [...]. [S. l.]: IEEE, 2010. p. 1-10.

SOUZA, P. H. G. F. de. *As causas imediatas do crescimento da renda, da redução da desigualdade e da queda da extrema pobreza na Bahia, no Nordeste e no Brasil entre 2003 e 2011*. Brasília: IPEA, mar. 2013. 23 p. (Texto para discussão, 1816).

TANTITHAMTHAVORN, C. *et al*. An empirical comparison of model validation techniques for defect prediction models. *IEEE Transactions on Software Engineering*, [s. l.], v. 43, n. 1, p. 1-18, 2016.

VIANA, I. Azevedo A. Vitelli V.;Organizadora; KAWAUCHI, Mary M.Organizadora; BARBOSA, Thiago T. Varanda V. (org.)Organizador. *Bolsa Família 15 Anos (2003-2018)*. Brasília: Enap, 2018.

ZAHARIA, M. *et al*. Apache Spark: a unified engine for big data processing. *Communications of the ACM*, [s. l.], v. 59, n. 11, p. 56–65, 2016.

Resumo

A obtenção de integração e uniformidade de dados provenientes de fontes heterogêneas é uma tarefa complexa que requer um planejamento estruturado para a recepção e tratamento de dados, sobretudo quando há grande volume de dados produzidos diariamente. Atualmente, existem alguns sistemas propostos em ambiente acadêmico para tal tarefa, sendo comum encontrá-los na indústria, porém sem apresentação de suas características, como seria de se esperar. Neste trabalho, propõe-se a concepção de uma arquitetura com solução para integração de dados oriundos de fontes heterogêneas. Para tanto, esta pesquisa teve como objetivo validar uma aplicação na qual dados extraídos de diversas tecnologias são direcionados a um ponto central de armazenamento e processamento de informações de forma homogênea, o que permite, em última instância, a apresentação dos dados de maneira padronizada. A ferramenta aqui proposta realiza, em tempo real, as principais etapas do processo de normalização de dados ao disponibilizar uma solução que promove integração de dados com o intuito de obter um gerenciamento operacional de ativos e pessoas. Tal sistema está centrado em georreferenciamento de dados.
Palavras-chave: Ciência de Dados. Eficiência. Integração. Fontes Heterogêneas. Arquitetura de Dados.

Abstract

Achieving integration and data uniformity from heterogeneous sources is a complex task that requires structured planning for the reception and treatment of data, especially when there is a large volume of data produced daily. Currently, there are some systems proposed in an academic environment for such a task, and it is common to find them in the industry, but without presenting their characteristics, as would be expected. In this work, we propose the design, the development process and the initial operational results of a solution for integrating data from historical sources. Therefore, this research aimed to validate an application in which the data extracted from several heterogeneous technologies are directed to a central point of storage and processing of information in a homogeneous way, which ultimately allows the presentation of data in a standardized way. The tool proposed here performs in real time the main stages of the data normalization process by providing a solution that promotes data integration in order to obtain an operational management of assets and people. Such a system is centered on data georeferencing.
Keywords: *Data Science. Integration. Heterogeneous Sources. Data Architecture.*

Uma arquitetura para integração de sistemas com fontes heterogêneas e big data

NELCI GOMES LIMA

Pós-graduada em Ciência de Dados e Big Data, pela Universidade Federal da Bahia (UFBA). nelcisgomes@gmail.com

LUCIANO REBOUÇAS OLIVEIRA

Doutorado em Engenharia Elétrica e de Computadores, pela Universidade de Coimbra (UC) e mestre em Mecatrônica, pela Universidade Federal da Bahia (UFBA). Coordenador do Intelligent Vision Research Lab, professor associado do Departamento de Ciência da Computação, Instituto de Matemática e Estatística, da UFBA. lrebouca@ufba.br

COM A EVOLUÇÃO das tecnologias computacionais, promoveu-se uma expansão no compartilhamento de dados. No que lhes concerne, os sistemas que geram dados tornaram-se indispensáveis à competitividade em vários setores empresariais, facilitando a comunicação e a tomada de decisões para melhoramento da visibilidade, eficiência, confiabilidade e segurança nos gerenciamentos operacionais. Isto é, um processo de coleta e envio de dados ligado a uma API que irá permitir a entrada de informações e, a partir de um modelo, converter dados de forma padronizada, mesmo que essas informações tenham sido originárias de fontes distintas.

Com base em levantamento bibliográfico, constatou-se que há recorrentes trabalhos relacionados à integração de dados. Por exemplo, o trabalho de Pinho, Boaventura-Cunha e Morais (2015), que diz respeito a técnicas de monitorização remota, redes de sensores, inspeção visual e integração de dados. No entanto, trabalhos com uso de tecnologias diversas ainda são poucos, isso

Esses aplicativos são projetados para detectar e reagir a alterações ou ocorrências de eventos conforme acontecem, no menor tempo possível

por se tratar de técnicas recentemente elaboradas em algumas áreas. Visando preencher essa lacuna, portanto, tem-se a justificativa deste trabalho, sobretudo pela necessidade de combinar dados de fontes heterogêneas, garantindo, assim, a qualidade deles, de modo a manter as informações sempre organizadas e simplificadas. A partir disso, tem-se a vantagem de dados padronizados e consistentes, que podem ser apresentados aos clientes de forma rápida.

A depender do cliente, a solução customizada também oferece uma tecnologia de dispositivos embutidos, sendo possível também a integração de dados originados de dispositivos outros, tais como sensores ou dispositivos sem fio. O processo de conexão e integração das diversas tecnologias é, geralmente, similar às conexões de APIs proprietárias e, muitas vezes, a informação obtida é pontual. Com base nos módulos de funcionamento, um diagrama de blocos foi construído com o objetivo de representar os diferentes componentes necessários para o desenvolvimento do projeto, que envolveu o uso das informações de rotinas geradas, utilizando-se a plataforma.

DO OBJETIVO

No presente artigo, apresenta-se a concepção de uma arquitetura de gerenciamento de dados provenientes de múltiplas fontes, tais como rádios digitais, *smartphones, tablets, beacons, tags* e controles de acesso, as quais podem ser acessadas através de uma conexão com *internet*, por meio de *clusters* de servidores hospedados na *cloud.* Segundo Bhadauria; Chaki (2011), essa é uma tendência que tem sido vista no cenário atual em quase todas as organizações. As vantagens de usar a computação em nuvem são: i) redução de hardware e custo de manutenção, ii) acessibilidade em todo o mundo e, iii) flexibilidade (processo altamente automatizado em que o cliente não precisa se preocupar com atualização de software). Para Zimmerle (2019), o processamento de eventos complexos e linguagens reativas são exemplos de soluções para as chamadas aplicações reativas. Esses aplicativos são projetados para detectar e reagir a alterações ou ocorrências de eventos conforme acontecem, no menor tempo possível. Todo esse processo ocorre através de sistemas ciberfísicos que enviam informações de um dispositivo a outro, ou a um sistema central, ou uma rede de dados. Os dispositivos estão interligados por algoritmos padronizados para interpretação que, de forma escalável, possibilitam a implementação contínua e otimizada da plataforma de soluções modulares, customizadas e georreferenciadas.

Borja (2014), afirma que a popularização de dispositivos com capacidade de comunicação tem levado ao iminente surgimento de uma

chamada internet das coisas. Neste contexto, a coleta de dados urbanos tem permitido o surgimento das cidades inteligentes, que necessitam integrar uma miríade de sistemas e dispositivos para o monitoramento urbano, possibilitando perceber informações relevantes. Este artigo propõe um modelo de arquitetura baseada em um barramento de serviços e eventos que permita o rápido desenvolvimento de aplicações.

Os pontos que servem de âncora para vincular e melhorar a capacidade de relacionar os dados são os de buscar informações relacionadas a georreferenciamento

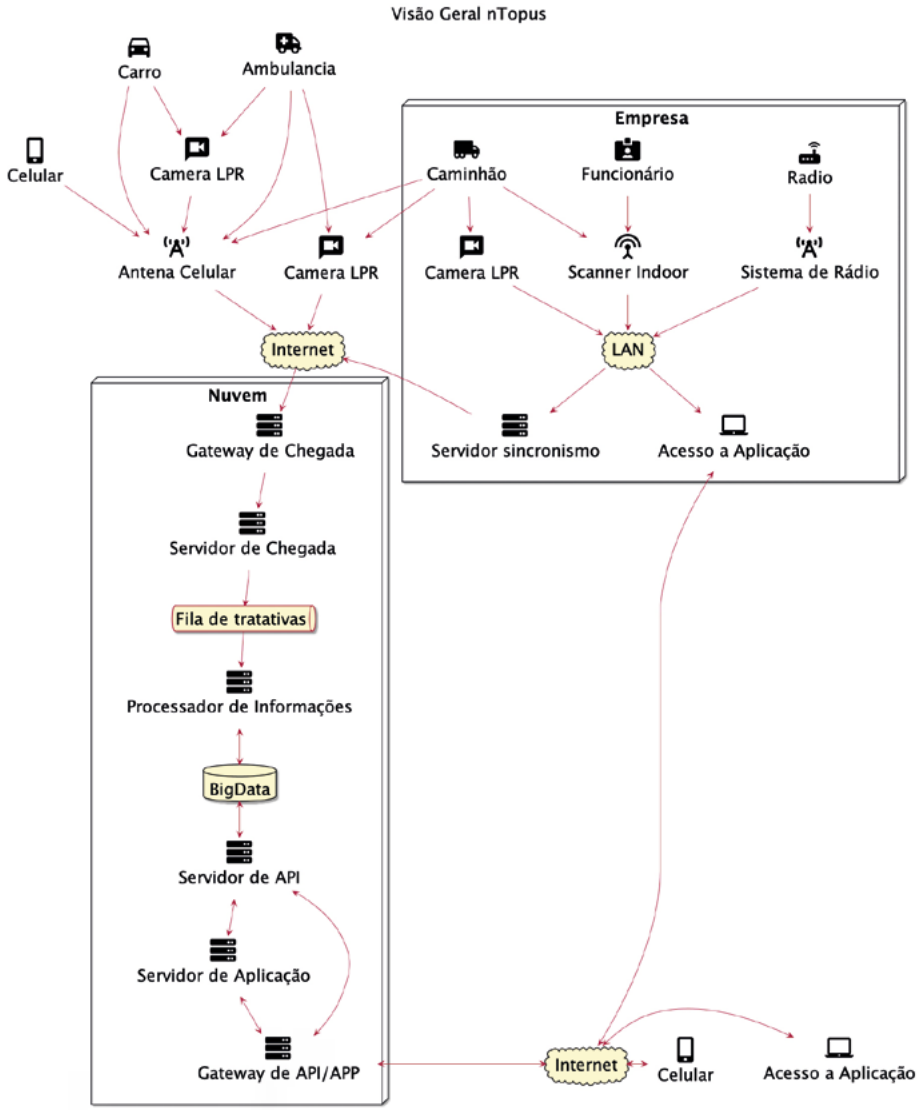
Na integração de dados, os dispositivos conectados aos sistemas podem ser direcionados à geração de alerta de condições ambientais nas quais trabalhadores estão inseridos, visando um gerenciamento operacional integrado. Nesse processo, o gerenciamento operacional de ativos e pessoas combina informações de localização por GPS, por *bluetooth* e por operacionais gerados a partir da integração e padronização de dispositivos móveis de diferentes tecnologias.

De acordo com Molina (2008), a sincronização de dados com base em descrições semânticas e *frameworks* de mapeamento de dados, com vistas a obter, como resultado, uma arquitetura que simplifica a integração de dados de fontes heterogêneas de maneira que minimiza o custo de sincronização. Com o uso de uma estrutura Big Data, a solução consegue aproveitar todo o volume de informações geradas no processo e, com isso, extrair conhecimento, como padrões de produção, melhorias nos produtos, aperfeiçoamento das máquinas e redução de custos.

Na Figura 1, apresenta-se uma visão geral da arquitetura com os componentes e processos que compõem o modelo de implementação seguido neste artigo. A seguir, mais detalhes.

Apesar de a ideia central do estudo ser lidar com fontes de dados heterogêneos, em sua essência, ainda se têm elementos comuns. Os pontos que servem de âncora para vincular e melhorar a capacidade de relacionar os dados são os de buscar informações relacionadas a georreferenciamento. Nos casos em que o georreferenciamento não é viável – seja por questões físicas, seja por incapacidade momentânea do dispositivo – a depender do dispositivo originador, podem existir informações relevantes e complementares para análises ou processamento posterior. Nesses casos o mecanismo utilizado é de marcar a informação recebida como georreferenciadamente rejeitada. Considerando a importância do tempo no projeto, um dos pré-requisitos para um funcionamento adequado são os elementos que estão no ponto de marcação de tempo e ao longo de todo processo estes devem estar com seus relógios sincronizados.

Figura 1
Visão geral da solução, a arquitetura e seus componentes para integração de dados



Fonte: Elaborado pelos autores.

Para os que têm acesso à rede GPS, isso é natural, porque para conseguir executar o algoritmo de triangulação isso faz-se necessário. Já os que não têm esse recurso precisam utilizar algum algoritmo de sincronismo de tempo, como NTP, buscando o servidor com menor latência possível e com um fluxo de verificação de divergência bem rígido. Inicialmente, ocorre a captura dos dados, responsável por saber lidar com as especificidades da tecnologia em questão. Uma vez que a informação foi compreendida pela coleta, ela é convertida para um padrão, o qual é genérico dentro do grupo de tecnologia geradora de dados rastreadores veiculares, rádios digitais, *beacons*, *bluetooths* etc.

Uma vez neste padrão, ela é então reconhecida para ser enviada para processamento, respeitando um padrão mais genérico no qual todos os pontos de coleta precisam respeitar para permitir um processamento uniforme dos dados. Quando o dado chega ao ponto de processamento, iniciam-se processamentos paralelos e sequenciais. Uns filtrando dados para saber se avançam para etapas seguintes, outros adicionando informações novas, e outros classificando o dado como não confiável, em caso de ocorrência. Uma vez armazenado, todo dado que chega é tratado como somente leitura, ou seja, só se adiciona informação, nunca se altera. Dentro do conceito, as informações podem abastecer bancos de dados diferentes, de modo a permitir a manipulação das informações no que se refere à sua natureza.

O sistema proposto prevê que, além de uma *interface* para entrada da consulta desejada, tenha-se opção de visualização dos dados, mesmo sem possuir conhecimento

Todo esse mecanismo abre espaço para utilização de pontos de gatilhos com regras configuráveis, permitindo tomar ações para bloquear um veículo, enviar uma mensagem para um rádio, enviar um e-mail e bloquear uma porta. Isto se torna possível graças ao modelo proposto de considerar cada ponto de processamento, sendo associado a um evento e, ao mesmo tempo, podendo gerar novos eventos. Isso permite uma flexibilidade para lidar com possíveis mudanças de como processar os dados e como atuar neles.

O final desse ciclo acontece nos pontos de acesso aos dados, considerando controles de usuário e perfis de acesso. É possível consultar os dados e ainda os receber à medida que eles acontecem. Os pontos de apresentação das informações ficam facilitados de buscar e lidar com isso devido ao processo de padronização feito na coleta bem como ao longo do processo. Isso permite uma facilidade na montagem de relatórios e telas de visualização, estáticas ou dinâmicas, já que há como ficar recebendo atualizações à medida que elas chegam e são manipuladas.

Este estudo não recomenda tecnologias para utilização nas etapas descritas anteriormente. Somente, quando for pertinente, pode-se citar alguma utilizada por questões de proficiência no processo do protótipo que demonstra a arquitetura. Com isso, o intuito é demonstrar quais tecnologias utilizar sem necessariamente entrar em detalhes de estado-da-arte para os componentes que compõem a solução, tais como:

- Integração com câmeras via RSTP OU ONVIF;

- Mapa de visualização 2D para localização *indoor*, na qual a consulta possa ser realizada corretamente e com facilidade pelo usuário. O sistema proposto prevê que, além de uma *interface* para entrada da consulta desejada, tenha-se opção de visualização dos dados, mesmo sem possuir conhecimento.

Apesar do amplo interesse em dados não estruturados e semiestruturados, os dados estruturados ainda representam uma porcentagem significativa de informações sob gerenciamento para a maioria das organizações

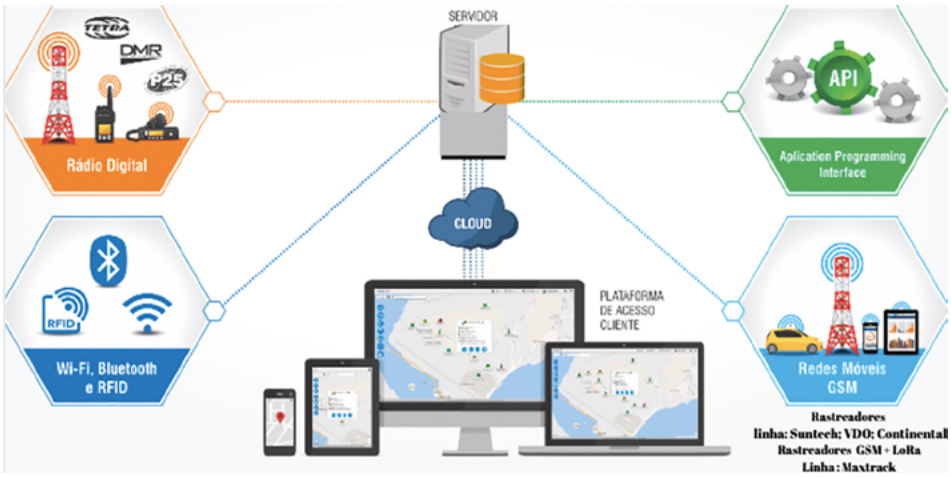
ARQUITETURA PROPOSTA

Para compreensão da complexidade do sistema e de sua viabilidade técnica, faz-se necessário realizar um levantamento dos principais conceitos que orientam sua idealização, construção e funcionamento. Dessa forma, apresenta-se aqui a arquitetura proposta buscando oferecer uma visão do funcionamento de todo o sistema em cada uma das suas fases. As principais formas de análise de dados presumem uma fonte de dados normalizada e estruturada, geralmente no formato de banco de dados. Em comparação com as arquiteturas tradicionais de conjunto de Big Data, normalmente há inclusão de dados massivos, sem estrutura fixa e com necessidades de análise em tempo real. A proposta desta solução trabalha exatamente com essa característica. Visando a necessidade de armazenar e processar tais conjuntos de dados, trazem-se desafios para as tecnologias já produzidas. Direcionadas para ambientes convencionais, a integração necessita de mecanismos confiáveis e procedimentos integrados para assegurar a consistência dos dados corporativos. Esses volumes excessivos demandam soluções que considerem questões de heterogeneidade e que possam se transformar em ambientes uniformes, com escalabilidade, tempo real, rapidez e privacidade. Uma aplicação para integração de dados possui vários ambientes.

Na Figura 2, apresenta-se a solução separada em blocos, os quais são combinados a aspectos de bancos de dados na arquitetura proposta, sistemas distribuídos e aplicações atuais do ponto de vista de serviços e (sistemas) *web*. Na simulação, todas as abordagens são executadas para funcionamento da aplicação de integração de dados. Diferentemente de dados coletados advindos das redes sociais e de comportamento cultural de pessoas, este estudo gera dados em tempo real de situações vividas diariamente no mundo corporativo através das atividades de seus colaboradores.

Para Quinto (2018), apesar de toda a empolgação com Big Data, a maioria dos dados críticos ainda é armazenada nos sistemas de gerenciamento de banco de dados relacional. Esse fato é suportado por estudos atuais on-line e confirmado pela experiência profissional do autor em vários projetos de Big Data e *business intelligence*. Apesar do amplo interesse em dados não estruturados e semiestruturados, os dados estruturados ainda representam uma porcentagem significativa de informações sob gerenciamento para a maioria das organizações, desde as maiores empresas e órgãos governamentais até pequenas empresas e iniciantes em tecnologia.

Figura 2
Tecnologias e ferramentas utilizadas na plataforma

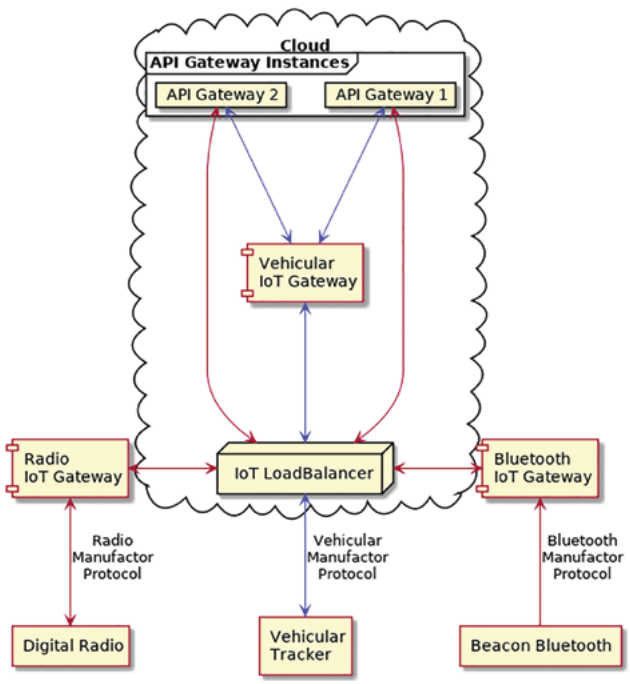


Fonte: NTOPUS Possibilidades (2020).

Captura dos dados

A Figura 3 ilustra toda a arquitetura de integração a partir da captura dos dados com os componentes e processos que compõem o modelo de implementação seguido neste artigo.

Figura 3
Diagrama com sistema de coleta de dados através de uma API



Fonte: Elaborado pelos autores.

Com a implantação de sensores de distância nos automóveis, torna-se possível localizar objetos próximos ou coletar dados sobre o próprio veículo

De modo geral, esses equipamentos não retêm informação por muito tempo e seu modo de funcionamento pode reportar a informação quando ocorre, ou somente quando consultados. Por exemplo, os rastreadores veiculares que reportam alerta ao moverem-se quando estão desligados, desenvolvendo o mesmo comportamento quando ligados. Alguns dispositivos são bem simples, a ponto de só conseguirem se anunciar para o ambiente, como é o caso dos *beacons bluetooth* (BLE).

Até o momento em que este estudo foi desenvolvido, conseguimos lidar com os seguintes equipamentos que servem para captura de dados:

- Rastreadores veiculares: (*i*) com GPS, percepção de ignição e botão de pânico; (*ii*) com GPS, percepção de ignição, botão de pânico, bloqueio de combustível, pinos de entrada e saída; (*iii*) com GPS, percepção de ignição, botão de pânico, bloqueio de combustível, pinos de entrada, saída e identificação de motorista;

- Rádios digitais: (*i*) Tecnologia Treta (Multi Fabricantes); (*ii*) somente envio de coordenada; (*iii*) envio de coordenada, códigos personalizados em botões pré-programados; (*iv*) envio de coordenada, códigos personalizados em botões e no teclado numérico, enviar e receber mensagens de texto;

- Tecnologia DMR (somente Motorola MOTOTRBO): *(i)* somente envio de coordenada; *(ii)* envio de coordenada e códigos personalizados em botões pré-programados; *(iii)* envio de coordenada, códigos personalizados em botões e no teclado numérico, enviar e receber mensagens de texto;

- Beacons detectados em ambiente fechado: *(i)* Beacons simples (só se anuncia no ambiente); *(ii)* Beacons com botão de ação (pode enviar um código personalizado);

- Beacons com sensores: *(i)* temperatura; *(ii)* umidade; *(iii)* pressão; *(iv)* acelerômetro;

- Câmera com tecnologia de detecção de padrões.

Dentro do escopo da visão computacional, pode-se utilizar aplicações focalizadas em identificar veículos através da placa (LPR). Outrossim, com a implantação de sensores de distância nos automóveis, torna-se possível localizar objetos próximos ou coletar dados sobre o próprio veículo. No entanto, a implementação da arquitetura requer a capacidade de enviar as informações coletadas/identificadas e a garantia da entrega da informação ao seu destinatário.

Alguns dispositivos têm condições de enviar os dados diretamente pela internet, uns somente em uma rede interna controlada, outros apenas no ambiente à sua volta, sem um destino. Essas diferenças fizeram com que a coleta se tornasse um dos pontos mais críticos e delicados para o projeto, exigindo arquiteturas diferentes para lidar com as situações já descritas. O módulo pensado para resolver essa questão foi o *IoT Gateway,* responsável por lidar com o protocolo distinto de cada tecnologia. Cada equipamento utilizado tem seu próprio protocolo, além de em alguns casos haver variação entre modelos, mesmo sendo do mesmo fabricante.

Alguns dispositivos têm condições de enviar os dados diretamente pela internet, uns somente em uma rede interna controlada, outros apenas no ambiente à sua volta, sem um destino

- Tipos de *IoT Gateways* criados para lidar com as situações: *I.* Com dados suficientes para enviar para processamento; *II.* Passivo: fica esperando conexão dos dispositivos; *III.* Conexão ativa: precisa conectar-se a um sistema já existente e cadastrar-se para receber dados; *IV.* Ativo: busca pelas informações dos dispositivos que estão ao seu alcance; *V.* Sem dados suficientes para enviar para processamento; *VI.* Injetores de dados georreferenciados

Nesse processo, é muito comum receber dados incompletos, o que demanda a adição de informações baseadas em cadastros prévios para poderem encaminhar os dados para processamento. Isto sem alterar em detalhes o modo como os protocolos de cada *IoT Gateway* funciona porque, após a informação ser decodificada, essa passa por um processo de conversão padronizado.

A título de exemplificação, podemos considerar que existe um conjunto básico de informações: versão dos dados, data e hora, código único de referência gerado para cada *gateway*, dispositivo originador, tipo *gateway* originador. Referentes ao georreferenciamento temos latitude, longitude, velocidade, precisão da coordenada, dentre outros. Considerando modelos de dispositivos, seguem informações adicionais.

- Rastreadores veiculares: *(i)* IOs: em forma de lista, identificando seu índice de referência, valor lido ou gravado, finalidade de ignição, bloqueio, trava elétrica, etc.; *(ii)* Motivo do envio: limite de tempo para envio de localização, distância excedida para envio de localização, veículo sendo rebocado, etc.; *(iii)* Hodômetro; *(iv)* Horímetro; *(v)* Nível de bateria do veículo; *(vi)* Nível de bateria reserva do rastreador;

- Rádios digitais: *i)* IOs: em forma de lista, identificando seu índice de referência, valor lido ou gravado; *ii)* Motivo de envio: limite de tempo para envio de localização, distância para envio de localização, retirado do carregador, acabou de ligar, acabou de desligar, trocou de canal de conversação de voz etc.; *iii)* Ponto

de transmissão/recepção filiada (equivalente às torres de celular para obtenção de cobertura); *iv)* Canal de voz atual;

- *Beacon* detectado em ambiente fechado: *i)* IOs: em forma de lista, identificando seu índice de referência, valor lido; *ii)* Local (ex. UFBA); *iii)* Prédio (ex. Instituto de Matemática); *iv)* Sala (ex. Laboratório do Lasid); *v*) MAC bluetooth do scanner detector do Beacon; *vi*) Nível da bateria reserva do Beacon; *vii*) Nível de sinal BLE lido pelo scanner.

É importante destacar que por se tratar de uma solução que depende de conexão de internet para envio dos dados para processamento, é de fundamental importância considerar que falhas de comunicação possam ocorrer momentaneamente, necessitando, por isso, que acumulemos os dados. Nesse sentido, um problema que pode ocorrer é o enfileiramento persistente, provenientes de, por exemplo, cenários de falhas no *IoT Gateway*.

Os dados deste último podem ser acumulados, mas ainda não sincronizados. A solução para esse problema foi utilizar um serviço de fila com persistência. Sendo assim, quando a informação chegou e foi convertida para o padrão interno, a mensagem foi colocada em uma fila de sincronismo na qual, ao se ter comunicação com a plataforma que está na nuvem, os dados começam a ser consumidos desta fila.

Uma vez que haja dados na fila de sincronização e o serviço que recebe os dados para processamento esteja apto para recebimento dos dados, as informações sofrem um processo de arrumação e padronização de unidades de medida

Uma vez que haja dados na fila de sincronização e o serviço que recebe os dados para processamento esteja apto para recebimento dos dados, as informações sofrem um processo de arrumação e padronização de unidades de medida. Há um padrão no qual os dados devem ser enviados para processamento, colocando todas as informações conhecidas referentes ao georreferenciamento com nomes padronizados, e tudo que foge a esse padrão vai para um campo “extras”, como nível de bateria.

O processo de sincronismo dos dados sempre respeita a ordem na qual os dados foram coletados. No caso de ficarem acumulados, são as informações mais antigas que se sincronizam primeiro. Para proteger os dados, o módulo de processamento funciona com uma comunicação *https*, de modo a criptografar os dados sincronizados. Outro ponto é que somente os *IoT Gateways,* com mecanismos de autenticação válidos (ex. *token http, oauth, etc*), conseguem enviar informações para a plataforma. Por se tratar de uma solução que por natureza não deve ficar expondo seus *IoT Gateways* para a Internet sem necessidade, fica a cargo de cada um “perguntar” ao módulo de processamento se existe algum comando para ser executado por algum dispositivo. Um bom exemplo de comando é o envio de mensagem da plataforma na nuvem para um

dispositivo. Por se tratar de uma solução na qual os equipamentos permitem algumas garantias no caso da mensagem temos confirmação de entrega e confirmação de leitura no dispositivo, que são informados a medida que ocorrem.

As principais arquiteturas para esses ambientes incluem clusters de processamento paralelo, através da utilização de bancos de dados tradicionais, multimodelos e plataformas de computação baseadas em memória

Processamento de dados

Na fase de processamento, as grandes quantidades de dados que são gerados a partir de fontes variadas impõem exigências distintas de armazenamento e processamento. Isso leva à apresentação de um grande desafio a estruturas de tecnologias computacionais convencionais, porque grandes volumes de dados requerem armazenamento que seja escalável e que tenha uma abordagem distribuída para possibilitar a consulta dessas informações de forma prática. As principais arquiteturas para esses ambientes incluem *clusters d*e processamento paralelo, através da utilização de bancos de dados tradicionais, multimodelos e plataformas de computação baseadas em memória. Na solução apresentada, as características de dados confiáveis são um diferencial, as quais são sempre aproveitadas e transformadas em dados significativos. Essa estruturação dos dados permite armazenar informações mais complexas, além de viabilizar fácil migração e replicação de dados.

Na aplicação apresentada, utilizamos *grafos*. Os bancos de dados baseados em *grafos* (especialistas em associação de um ou mais dados entre si) são indicados para sistemas que necessitam de relacionamento de dados para prover informações para o usuário final, tais como redes sociais, sistemas de recomendação e serviços orientados por localização.

A plataforma precisa naturalmente ter um conjunto de cadastros prévios: *i)* IoT Gateways; *ii)* Dispositivos e *iii)* Sistemas. Esses cadastros são consultados à medida que novas informações são recebidas. Dependendo de como está o cadastro, a informação pode ser descartada.

Chegada dos dados

De modo a permitir que os *gateways* tenham a maior possibilidade de enviar os dados e também de modo a deixar que estes dados possam estar disponíveis para processamento, fazemos uso novamente do sistema de filas persistentes. Uma vez que a aplicação confirma o armazenamento na fila, essa devolve um "OK" para o *gateway.* Em caso de falha nesse processo, é reportado um sinal de "Error" e o *gateway* reenvia esse mesmo dado até que ele possa ser armazenado na fila para processamento. Outro ponto importante, na chegada dos dados, é que algumas consultas são feitas antes de aceitar o armazenamento dos dados na fila, como *i) token* de acesso do IoT *Gateway* e *ii)* se o sistema é conhecido pela aplicação.

Tudo que é consumido das filas de chegada é tratado como um evento de chegada de dado de localização, o qual pode gerar novos eventos

A não existência do dispositivo não impede a chegada de dados. No entanto, a incoerência na versão do dado que está sendo sincronizado pode afetar e impedir a chegada dos mesmos, seja por uma versão descontinuada, seja por uma versão ainda não interpretável.

Dinâmica de eventos

A partir desse ponto, tudo que é consumido das filas de chegada é tratado como um evento de chegada de dado de localização, o qual pode gerar novos eventos. Para uma solução mais flexível e escalável foi constatado que uma abordagem baseada em eventos favorece o adicionamento de novos comportamentos à aplicação, isso sem precisar ficar mudando códigos já funcionais e consolidados. Outro ponto positivo desta abordagem é a complexidade de lidar com cada especificidade de forma pontual e assertiva. Mudanças de comportamento, na sua essência, passam a ser criação de novos tratadores de eventos, com transferência de tratadores de eventos antigos para outros processadores, sem necessariamente reescrever seu código.

Outro ponto de destaque é a boa prática de criar evento por padrão para cada tratador, independentemente de já ter (ou não) tratativa para o resultado que foi abordado. Além de facilitar no fluxo de mudanças comportamentais da aplicação, há a continuação do fluxo com novas necessidades e especificidades, tornando-se trivial e relativamente seguro de implementação, com impacto e risco baixos na mudança do comportamento atual da solução.

Tratadores

De modo a facilitar o entendimento do fluxo de leitura durante o processo, foram criados níveis de tratadores, os quais são separados por operação. Os “Nativos” lidam com tudo que é originado de uma ação direta na plataforma, já os “derivados” são oriundos dos eventos “nativos”. Um fluxo básico na aplicação consiste em lidar com uma sequência de tratadores, com seus respectivos eventos a partir de uma localização.

- Evento de chegada de localização */Native. Location. Receivea: i)* tratador de armazenamento no Log de Localizações – armazena a localização no banco de dados; *ii)* gera novo evento de Nova Localização Armazenada /Derivate. Location. Created.

- Evento de Nova Localização Armazenada */Derivative. Location. Created: i)* tratador de tradução de coordenada em endereço – consulta serviço de mapas e pede atualização no Banco de Log de Localização para anexar a informação de endereço; *ii)* gera evento de identificação de motorista – consulta o serviço de mo-

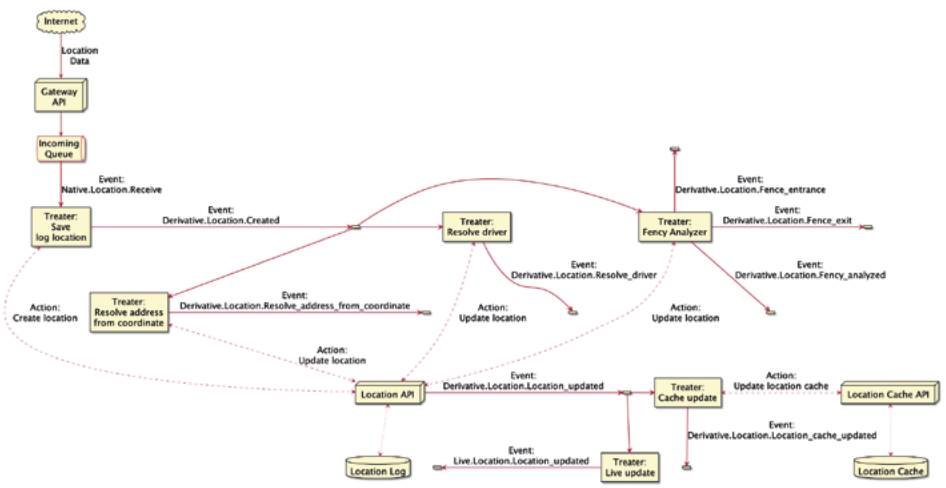
toristas de veículos e pede atualização no Banco de Log de localização para anexar a informação do motorista.

- Evento de motorista identificado /Derivative. Location. Driver discovered: *i)* tratador de análise de cercas – identifica com quais cercas o dispositivo está no momento e pede atualização no Banco de Log de localização para anexar a informação de cercas, incluindo quais são novas e quais não estão mais; *ii)* gera evento de entrada de novas Derivate. Location. Fence entrace; *iii)* gera evento de saída de cerca para as que não estão mais */Deriva-tive. Location. Fence exit.*

- Evento de localização atualizada – *Derivate. Location. Updated.* – Gerado quando alguma informação é anexada ao Log de localizações: *i)* tratador de atualização de memória transitória – atualiza o banco de dados de *cache* com a informação mais nova da localização deste dispositivo; *ii)* gera evento de *cache* atualizado /Derivate. Location. Cache Updated; *iii)* tratador de informação ao vivo – propaga uma informação atualizada para um possível usuário visualizador conectado na plataforma; *iv)* gera evento de localização ao vivo */Live. Location. Updated.*

Na Figura 4, a seguir, apresenta-se o diagrama do fluxo baseado em Evento, isto é, uma sequência de tarefas nas quais o processo acontece. Por uma mensagem, indicador, notificação ou algo similar significa que uma ocorrência aconteceu e foi registrada. Aqui, permitiu-se descrever todos os passos dos processos no menor nível de granularidade.

Figura 4
Diagrama de fluxo baseado em eventos



Fonte: Elaborado pelos autores.

Existem abordagens nas quais o serviço de fila, além deste armazenamento, sequencial, permite também o funcionamento do modelo baseado em eventos

Pontos de armazenamento

Como é possível observar no fluxo anterior, temos alguns pontos de persistência de dados. Por isso, foram utilizadas algumas tecnologias para melhor lidar com essas questões, e, dessa forma, aproveitar o melhor de suas características. Nesse sentido, serviços de fila foram utilizados para dar persistência imutável e cronologicamente sequencial aos acontecimentos, permitindo manter uma sequência lógica nas narrativas. Um efeito colateral a essa abordagem é que – ao ter um tratador que não processe os dados em tempo hábil – a chegada dos dados gera acúmulo de informações nas filas.

Existem abordagens nas quais o serviço de fila, além deste armazenamento, sequencial, permite também o funcionamento do modelo baseado em eventos. Dessa forma, torna-se um ponto central no qual todos os tratadores podem se filiar e, ao mesmo tempo, gerar novos eventos sem a necessidade de uma intermediação ou mudança de código na extinção de eventos na solução. Tudo isso torna dinâmica a sua manutenção e expansão. Como o propósito de algumas etapas é adicionar informações correlatas aos dados no seu armazenamento, precisamos de dois bancos de dados já consolidados:

- Relacionais – responsáveis por guardar informações que não sofrem muitas alterações, além de serem em sua grande maioria mais consultadas para servir de complemento, dando sentido à informação que chegou ou foi gerada;

- Não relacionais – lidam com as informações que são muito alteradas ou que têm um grande volume sendo gerado, além de permitir um gerenciamento de sanitização mais simplificado, apesar do risco de perda de dados com políticas mal pensadas.

  As utilizadas no trabalho foram:

  1. cache: responsável por lidar com as informações do "agora", momento presente. Qualquer informação que mude muito e que precisa ser consulta de forma rápida fica armazenada nessa tecnologia. Por exemplo, a última localização recebida por um dispositivo; a última vez que um dispositivo falou com a plataforma, ou mesmo alerta ativo (ocorrendo no agora) em um mecanismo;

  2. Log/Dados históricos: informações que precisam ser registradas com peso de histórico e, ao mesmo tempo, com complementos de dados para facilitar a mineração dos dados em momentos posteriores são o forte desta tecnologia. Dados re-

ferentes às localizações recebidas ficam aqui, considerando-se as informações de contexto já anexadas, como detalhes dos dispositivos (apelido, com quem estava no momento, etc.), com tratativas externas (código de chamado de atendimento, endereço, etc.), dados de pessoas (nome do motorista ou equipe do veículo).

A organização na estrutura de codificação dos eventos permite facilmente que o módulo de escuta ao vivo vincule-se a qualquer evento para começar a receber qualquer informação quando ocorre

Live

É importante destacar que uma solução que se propõe a atuar no presente, propagar tornar-se crucial. Com isso, na abordagem apresentada anteriormente, podemos perceber que a arquitetura permite anexar tratadores de propagação de evento ao vivo a partir de qualquer ponta da cadeia de tratativas. A organização na estrutura de codificação dos eventos permite facilmente que o módulo de escuta ao vivo vincule-se a qualquer evento para começar a receber qualquer informação quando ocorre.

Processadores de eventos e ações

Outro propósito pensado na modelagem da solução foi o de conseguir executar uma ação virtualmente a partir de qualquer coisa que tenha acontecido. Novamente, a abordagem de eventos permite anexar um ouvinte de eventos para, com isso, aplicar uma regra de filtragem e disparar uma ação. Alguns exemplos para demonstrar o uso deste recurso:

1. enviar um e-mail informando que uma viatura excedeu o limite de velocidade dentro de uma região crítica na área da empresa;
    - evento: cerca analisada;
    - ação: envio de e-mail, usando informações de cercas nas quais o veículo se encontra;
2. bloquear uma viatura ao exceder os limites da área de trabalho;
    - evento: saída de cercas; filtrando com uso de informações de cercas nas quais o veículo deixou de estar;
    - ação: bloqueio de veículo;
3. enviar mensagem para o rádio de um agente ao perceber que ele está há muito tempo parado no mesmo ponto, solicitando-o que siga com sua ronda (considerando que existe um tratador de análise de dispositivo parado por muito tempo);

Exportações automatizadas também são tratadas como exibição ou visualização

- evento: dispositivo parado há muito tempo;
- ação: envio de mensagem para um dispositivo informando para seguir com a ronda.

Exibição dos dados

Podemos considerar como exibição qualquer tentativa de acessar as informações disponibilizadas pela solução. Tanto as informações já armazenadas quanto as que estão chegando ou sendo geradas ficam passíveis de disponibilização. Exportações automatizadas também são tratadas como exibição ou visualização. Para se ter uma estrutura confiável e segura, precisam ser considerados alguns componentes. Aspectos relacionados às técnicas de desenvolvimento consideradas seguras não entram no escopo deste trabalho, mas são fortemente recomendadas em seu uso. A arquitetura para visualização das informações considera algumas camadas principais, representado na Figura 5:

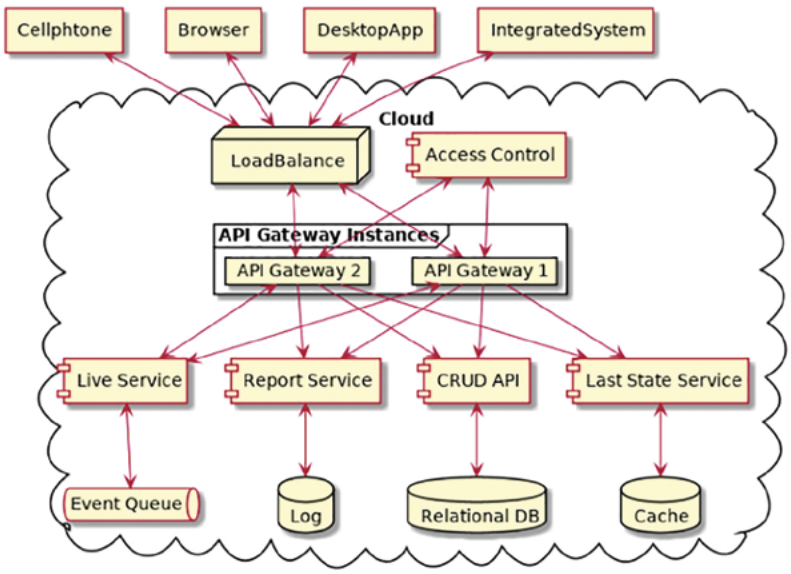
- Load Balance: responsável por permitir que as tentativas de acesso sejam espalhadas entre as instâncias disponíveis para processamento dos dados solicitados. Esse componente permite redimensionar toda a arquitetura para suportar os volumes de acessos e, ao mesmo tempo, reduzir a hipótese de uma negação de serviço;
- *API Gateway:* permite ter um portão principal no qual todos os acessos de um determinado conjunto de APIs está sendo disponibilizado. Essa abordagem permite simplificar implementações dentro da infraestrutura, deixando alinhada a essa camada uma integração com um sistema de controle de acesso e um controle de consumo da API por parte do cliente (limite de acessos por segundo);
- Serviço de autenticação e autorização: esse serviço permite que as camadas mais internas não precisem obrigatoriamente se preocupar com quem é e se pode ou não acessar a informação. Ela sabe identificar se um usuário está logado na plataforma e se é preciso adicionar alguma informação para a rota de acesso no momento;
- Serviços de *Report*: são serviços responsáveis por receber as solicitações de geração de relatórios para, então, devolvê-los aos seus clientes. Essa camada consegue gerar dados em diversos formatos, como CVS, PDF, JSON, XML etc.;

- Serviços de *Live*: esse serviço permite cadastramento para receber dados de praticamente qualquer evento gerado pela plataforma;

- API de cadastros: esse serviço permite lidar com consultas e operação de criação, edição, exclusão e listagem de informações de cadastro da plataforma;

- Serviço de informações recentes: esse serviço provê o estado mais atual das informações coletadas e processadas pela solução

Figura 5
Gráfico com estruturas de componentes para visualização



Fonte: Elaborado pelos autores.

Abastecer fontes de dados externos/complementares

Uma das formas mais simplificadas de visualização dos dados é a transferência para outro sistema no qual eles podem ser convertidos e interpretados dentro do universo do qual faz sentido. Nesta mesma arquitetura, pode-se fazer uma combinação de consulta de log com vinculação para recebimento de novos acontecimentos à medida que eles ocorrem.

Interface para carregar informações dos dados armazenados

Com fins de simplificação, a solução prevê meios para permitir a visualização dos dados. Uma *interface* de usuário foi criada para consumação desta estrutura previamente descrita. Nesta, pode ser feito um *login*

Uma preocupação primária na chegada dos dados é a de não impedir a chegada de dados de dispositivos que não estão cadastrados na plataforma

para se ter acesso à aplicação, permitindo acessar relatórios por período, filtrando por dispositivo, cercas e limite de velocidade. Outrossim, permite também escolher os campos no quais serão considerados na geração do relatório, permitindo esconder informações não relevantes para uma determinada situação.

Numa *interface,* é possível acompanhar em um mapa do Google Maps o local onde cada dispositivo está em determinado momento e qual era a última localização de um dispositivo em não-comunicação. Na mesma tela, é possível visualizar e desenhar cercas para serem utilizadas nas regras. Outro ponto importante desta tela é que ela faz uso do serviço de *Live* e de informações recentes. No momento de carregar a tela de mapa, utiliza-se da API de informações recentes e vincula com *Live.* A partir deste momento em diante, todo processo de atualização de informações é feito sob demanda (à medida que elas chegam pelo serviço de *Live*), isso permite baixo uso de banda de dados para acesso via *smartphone*.

Detalhes da arquitetura: preocupação na chegada dos dados

Uma preocupação primária na chegada dos dados é a de não impedir a chegada de dados de dispositivos que não estão cadastrados na plataforma. Considera-se aceitável do ponto de vista de esquecimento de cadastro não gerar perda de informação. Isso possibilita gerar também métricas de possíveis falhas de cadastro, além de ajudar na procura por dispositivos que estão usando a infraestrutura indevidamente, permitindo, assim, ir até sua localização. Para mais, dependendo do caso, pode-se tomar providências para correção da situação. Outro ponto crítico já citado anteriormente é quanto ao fluxo de dados recebidos ou gerados, os quais se confrontam com a capacidade de processamento dentro do ciclo de tratativas de eventos.

Graças ao desenho sugerido neste artigo, podemos lidar com essa situação – em grande parte – somente aumentando o número de tratadores para o mesmo evento que está executando a mesma tratativa. Um segundo benefício é que temos, devido à distribuição, uma visão facilmente coletável do desempenho computacional de cada tratador, permitindo-se, com isso, tomar conhecimento de quais são seus pontos de gargalo. Do mesmo modo, permite-se encontrar justificativas para substituição do tratador por uma versão mais eficiente, seja trocando linguagem, seja repensando seu algoritmo.

Preocupação na integração com outras aplicações

A preocupação na integração com outras aplicações é um ponto que exige necessária atenção, sobretudo quando a integração é a criação

de contas de aplicação, as quais, na prática, representam um acesso ao sistema sem um usuário identificado pela plataforma. Contas deste tipo são percebidas como uma conta de aplicação, deixando explícito que se trata de um acesso a um sistema externo não vinculado diretamente a um usuário. Geralmente, essa conta dá acesso a grupos de entidades dentro da aplicação, podendo atuar como um ou como mais de um usuário.

Outra preocupação com as integrações é a de facilitar sua utilização. Por esse motivo a arquitetura proposta permite que possa haver uma camada de tradução personalizável onde o ponto de acesso de informações pode mapear como a estrutura de dado pode ser gerada, desde o nome dos campos até seu formato. Em determinados cenários, dependendo da viabilidade, pode-se disponibilizar o acesso via exportação dos dados, direto para um banco de dados de preferência da aplicação que irá consumir da plataforma.

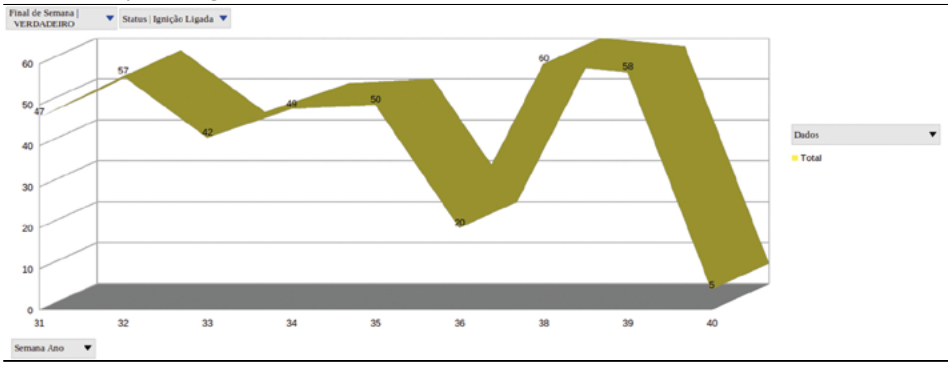
Dependendo da viabilidade, pode-se disponibilizar o acesso via exportação dos dados, direto para um banco de dados de preferência da aplicação que irá consumir da plataforma

ESTUDO DE CASO – ATUAÇÃO NO SETOR PÚBLICO

A arquitetura apresentada neste artigo se manteve dentro de um escopo mínimo. A plataforma que está sendo posta à prova tem recursos adicionais que tornaram seu uso bastante interessante para os órgãos futuramente citados.

Na Figura 6 ilustra uma análise simples através da utilização de Big Data, para gerenciamento de frotas. Nela, é possível monitorar o uso de dispositivos cadastrados na plataforma, acessar localização e verificar quando e quais motoristas estão em cada veículo. O gerenciamento de frotas com base em dados diários tem o objetivo de otimizar processos, reduzir custos e facilitar para os profissionais.

Figura 6
Análise simples de gerenciamento de frotas



Fonte: Elaborado pelos autores.

Um bom exemplo para compreender o mecanismo de funcionamento é notar que tanto Transalvador como SAMU e Guarda Municipal utilizam o mesmo rádio

No gráfico, o gestor tem a possibilidade de saber quais de seus ativos foram usados e, assim, obter informações sobre o planejamento, gerando padrões de uso, bem como os impactos do uso fora do horário de expediente e nos fins de semana.

É possível também marcar pontos de apoio ou referência, visualizar estado do trânsito direto no mapa e ter o estado operacional dos ativos monitorados. O estado operacional pode ser modificado, dependendo da tecnologia utilizada pelo próprio agente em campo, dando uma agilidade à operação. O mesmo equipamento pode gerar estados operacionais diferentes dependendo de como está configurado na plataforma para cada grupo de ativos. Uma boa forma de visualizar isso é com o rádio digital, que tem um teclado numérico que, quando pressionado, permite enviar um código que representa seu estado operacional.

Esse código pode ser traduzido para qualquer situação e essa tradução pode ser diferente dependendo de como a plataforma entende o originador do código. Um bom exemplo para compreender o mecanismo de funcionamento é notar que tanto Transalvador como SAMU e Guarda Municipal utilizam o mesmo rádio, porém se pressionarmos o botão "1" isso significa coisas diferentes, a depender do órgão e do grupo do ativo, pode significar "almoço" para um, "em atendimento" para outro e "intercorrência" em um terceiro.

Existe mais um recurso na plataforma, criado para facilitar lidar com comportamentos mais específicos e personalizá-los por cliente. Uma área de gestão de regras que permite trabalhar com os eventos que estão acontecendo e executar ações a partir disso, sendo envio de e-mail, escrever mensagem em grupo do Telegram, modificar um estado operacional de um ativo, enviar uma mensagem, sincronizar com um serviço externo etc.

A plataforma também dispõe de uma ferramenta para análise visual de um momento, chamado de linha do tempo. Esta permite ver uma foto interativa de como estavam seus ativos em um dado momento de um dia específico, permitindo avançar e voltar no tempo e vendo como estavam se comportando seus ativos. Este recurso serve para auxiliar em uma análise de sinistro, melhorar procedimentos ou estratégias de atendimento, além de permitir uma auditoria sobre se alguém não seguiu um procedimento, baseado em sua localização.

Outra boa forma de utilização da aplicação é através do sincronismo por faixa de hora. Um bom exemplo corresponde ao da Diretoria de Iluminação Pública, da Secretaria Municipal de Ordem Pública (Semop). Esta permitiu que seus técnicos de manutenção pudessem ser monitorados para saber como estavam suas atividades dentro dos limites de tempo.

Nesse contexto, considerando que um atendimento de substituição leva em média 30 minutos, quando um técnico informa que recebeu um serviço, o sistema consulta a plataforma na faixa de atendimento 30 minutos antes e depois. Se o técnico não esteve no local ou se ficou mais tempo do que o normal, essa informação segue para o setor de análise para que seja realizada identificação dos possíveis problemas no material de manutenção ou na saúde dos próprios postes (ou mesmo dos postes de iluminação). Assim, permite-se tomar medidas mais efetivas, as quais serão consideradas decisões de médio a longo prazos. Além disso, pode-se fiscalizar os técnicos quanto às informações equivocadas.

O uso do estado operacional permite que um ativo trocando de zona por estar em atendimento pode ser considerado normal, porém a mesma situação em modo ronda, pode ser considerada anômala

- Localização *on-line* baseada em grupos operacionais;
- Recursos de otimização de visualização *on-line*;
- Localização instantânea de usuário por nome ou por dispositivo de rastreamento com filtros de visualização instantânea por usuário, ou mesmo por grupo de usuários em três modos de visualização de mapas;
- Mapa, satélite ou *street view;*
- Histórico visual de trecho percorrido;
- Dados históricos completos para gerenciamento – linha do tempo (com este recurso, é possível avaliar de forma ampla a condição em um determinado período passado).

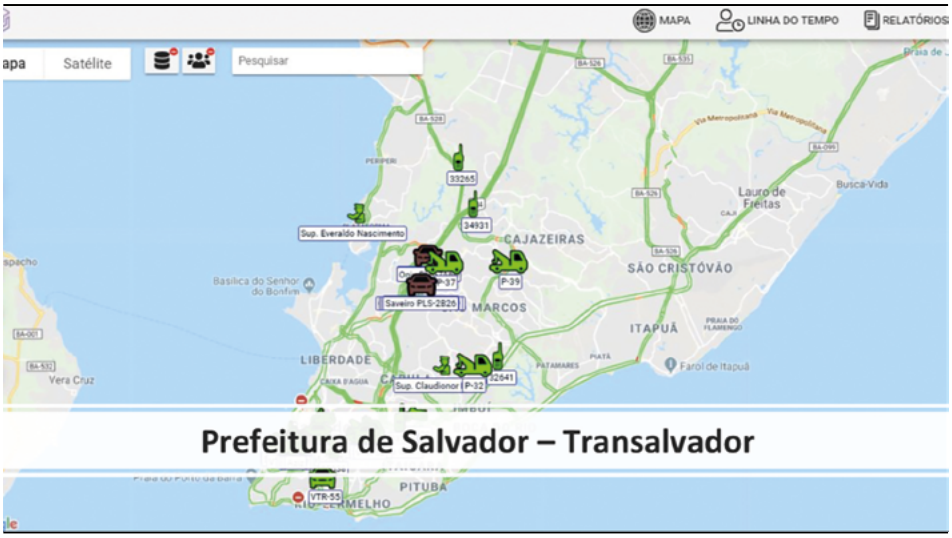
Transalvador – Superintendência de Trânsito do Salvador

Existem cercas cobrindo a cidade e, dentro dela, as zonas de ronda e atendimento, graças a todo mecanismo, já descrito anteriormente, a ferramenta se mostrou capaz de analisar um veículo atuando fora da sua zona ou da cidade e gerar uma mensagem via Telegram para um grupo com pessoas que tem condições de atuarem em tal situação. Nos casos de saída da cidade, a ação mais drástica, depois de avisos formais ou ciência da situação, é o bloqueio do veículo.

O uso do estado operacional permite que um ativo trocando de zona por estar em atendimento pode ser considerado normal, porém a mesma situação em modo ronda, pode ser considerada anômala.

O órgão tem uma ferramenta centralizada que gera toda a operação; um passo natural é compartilhar informações que ambas as plataformas não têm para tornar o processo de atendimento mais fluido para o

Figura 7
Painel de monitoramento Transalvador



Fonte: NTOPUS Possibilidades (2020).

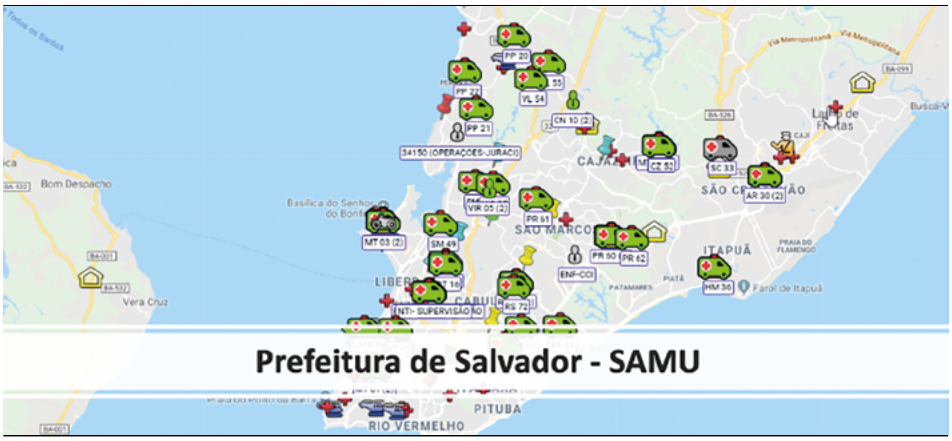
operador na central. Desde qual viatura está mais próxima, até atualizar automaticamente qual a situação de um atendimento (última localização conhecida, estado operacional etc.) ou reportar rapidamente uma emergência através dos recursos que o rádio utilizado pelos agentes permite, como, por exemplo, identificar queda ou desmaio.

A ferramenta de linha do tempo também é utilizada para investigar atendimentos não conformes.

SAMU – Serviço de Atendimento Móvel de Urgência

O uso da plataforma serve principalmente para auxiliar os médicos a saberem onde estão as ambulâncias que foram acionadas para o atendi-

Figura 8
Painel de monitoramento SAMU



Fonte: NTOPUS Possibilidades (2020).

mento. Além disso, o processo de encontrar a ambulância mais próxima para o atendimento também faz uso do estado operacional, permitindo saber qual veículo está realmente disponível.

A ferramenta de linha do tempo também é utilizada para investigar atendimentos não conformes, melhorar atendimento, repensando pontos de apoio, para reduzir o tempo de chegada.

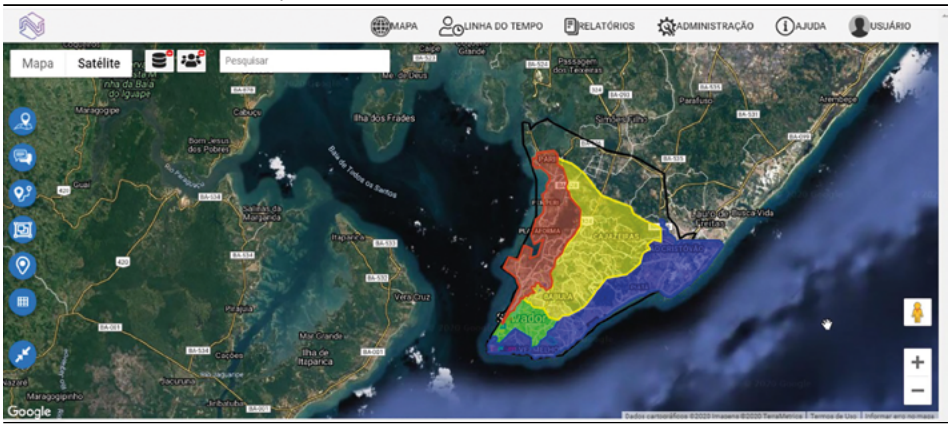
Guarda Municipal

Utiliza a solução basicamente para acompanhar a disposição dos grupamentos de acordo com a área operacional. Fazem os controles funcionais utilizando cercas eletrônicas, abaixo na Figura 9 um exemplo, no Carnaval de Salvador de 2020, foi utilizada a solução com rastreador veicular e rádio para desenvolvimento e monitoramento das atividades de patrulhamento preventivo que circulam nos circuitos.

As equipes, quando em campo, usam os status operacionais para acompanhar alguma ocorrência e/ou indicação de refeição, que seria o período de não movimento. A depender do alerta recebido os responsáveis pelo sistema podem acionar o bloqueio do veículo.

Os satélites enviam as coordenadas de posicionamento do carro através do módulo GPS integrado. As coordenadas são enviadas para uma central de processamento de dados através da telefonia móvel, que utiliza o canal de dados exemplificado na Figura 10.

Figura 9
São as cercas das bases operacionais de Salvador



Fonte: Kofre Telecomunicações (2020).

Figura 10
Painel de monitoramento da Guarda Municipal

Fonte: Kofre Telecomunicações (2020).

CONCLUSÃO

Com base no trabalho aqui descrito, foi possível concluir que os objetivos listados inicialmente para o projeto foram alcançados. A arquitetura para integração de sistemas com fontes heterogêneas foi desenvolvida de forma que possa ser expandida no futuro, adicionando novas funcionalidades aos sistemas criados a partir dela. A estrutura criada no projeto engloba de forma geral os principais pontos relacionados ao gerenciamento de recursos operacionais de algumas empresas em Salvador, hoje funcionando de forma dinâmica e com retornos eficientes.

Uma análise preliminar feita por membros da empresa concluiu que o *software* criado a partir da arquitetura proposta facilita a adição de novos recursos/serviços e sua modularização permite expandir para versões mais robustas. Os estudos de casos comprovam os seguintes benefícios:

- Rádio tetra em versão mais robusta (comunicando diretamente com sistema de processamento central da tecnologia, que varia por fabricante) e adição de novos *gateways* de tecnologia (sem mudança na aplicação de processamento);

- Rádio MOTOTRBO em versão mais robusta (comunicando direto com sistema de processamento central da tecnologia, que varia por arquitetura de implantação) e novos modelos de rastreadores sendo processados. Novas funcionalidades nos GW existentes (com baixo nível de alteração no modelo principal de dados);

- Acoplamento de dispositivos de identificação de pessoas aos rastreadores veiculares, permitindo saber quem está conduzindo o veículo;

- Envio de mensagens ao Telegram;

- Telas de acompanhamento em tempo real das atividades do colaborador e assim validação de novas formas de atuação no gerenciamento operacional com baixo impacto no processamento central;

- Espera de uma resposta vinda da integração com o Telegram e uso de dados minerados de cada área de atuação do gestor responsável, bem como o uso de informações para melhor atender cada empresa;

- Dispositivos físicos que permitem interação humana podem ser utilizados para ter uma comunicação rápida com elementos que estão no pós-processamento de forma transparente;

- A abordagem proposta traz inovações, através dos mecanismos de configuração física e lógica de tecnologias já usadas no dia-a-dia das empresas, permitindo construir sistemas que interajam de forma dinâmica. Pode ser integrada a um número variável de componentes e, cada componente, no que lhe concerne, pode oferecer uma variedade de operações e serviços de rastreamento, comunicação, coleta e armazenamento de dados. Esta flexibilidade torna mais ampla e personalizada a escolha e integração de componentes, bastando para isso escolher uma combinação adequada de serviços da empresa que irá usar o modelo proposto;

- Futuramente, com a consolidação dos modelos de dados de transmissão e armazenamento como modelos abertos (*open source*) na comunidade irá facilitar sua adesão e incentivar que novos modelos possam surgir;

- Espera-se camadas de montagem de rotas inteligentes para atividades de entrega/atendimento nas empresas em que o *software* atua.

A arquitetura proposta, apesar dos benefícios, traz dificuldade em garantir cenários mais realistas de teste de *software*. Isto adiciona um desafio a como gerir a infraestrutura e lidar com os mecanismos de escalar seus componentes, exigindo ter um bom monitoramento e log de todos os componentes e servidores que estão rodando a solução. Outro desafio que fica para futuras discussões é como controlar o acesso a aplicações externas e manter a segurança da plataforma sem perder sua flexibilidade.

Outro desafio que fica para futuras discussões é como controlar o acesso a aplicações externas e manter a segurança da plataforma sem perder sua flexibilidade

REFERÊNCIAS

ALAM, M. M.; HAMIDA, E. B. Strategies for optimal mac parameters tuning in ieee 802.15. 6 wearable wireless sensor networks. *Journal of Medical Systems*, [s. l.], v. 39, n. 9, p. 106, Aug. 2015. Disponível em: https://link.springer.com/article/10.1007/s10916-015-0277-4. Acesso em: 10 jun. 2020.

BHADAURIA, R.; CHAKI, R. *A survey on security issues in cloud computing*. [S. l.: s. n.], 2011. Disponível em: https://www.bibsonomy.org/bibtex/2ace3e17fbcf0af4cda343664736d4876/gizmoguy. acesso em: 5 ago. 2020.

BORJA, R.; GAMA, K. Middleware para Cidades Inteligentes baseado em um Barramento de Serviços. *In*: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO (SBSI), 10., 2014, Londrina. *Anais* [...]. Porto Alegre: Sociedade Brasileira de Computação, maio 2014. p. 584-590. Disponível em: https://sol.sbc.org.br/index.php/sbsi/article/view/6147. Acesso em: 10 jun. 2020.

DEGAN, J. O. C. *Integração de dados corporativos*: uma proposta de arquitetura baseada em serviços de dados. 2005. 96 f. Dissertação (Mestrado) - Universidade Estadual de Campinas, Instituto de Computação, Campinas, SP, 2005. Disponível em: http://www.repositorio.unicamp.br/handle/REPOSIP/276489. Acesso em: 5 ago. 2020.

KOFRE TECNOLOGIA. *Soluções em tecnologia de sistema de radiocomunicação*: tecnologia à serviço da vida. Disponível em: http://www.kofre.com.br. Acesso em: 22 abr. 2020.

MCAFEE, A. *et al*. Big data: the management revolution. *Harvard Business Review*, [s. l.], v. 90, n. 10, p. 60–68, 2012.

MOLINA, M. S. *Proposição de uma arquitetura de sincronização de fontes heterogêneas de dados utilizando frameworks de mapeamento de dados*. 2008. 170 f. (Trabalho de Conclusão de Curso de Graduação em Sistemas de Informação) - Universidade Federal de Santa Catarina, Florianópolis, 2008.

NTOPUS POSSIBILIDADES. *Sistema de integração multiplataforma para gestão operacional*. Disponível em: https://www.ntopus.com.br. Acesso em: 11 ago. 2020.

PINHO, T.; BOAVENTURA-CUNHA, J.; MORAIS, R. Tecnologias da eletrónica e da computação na recolha e integração de dados em agricultura de precisão. *Revista de Ciências Agrárias*, Lisboa , v. 38, n. 3, p. 291-304, set. 2015. Disponível em: http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S0871-018X2015000300003&lng=pt&nrm=iso. Acesso em: 29 ago. 2020.

QUINTO, B. *Next-Generation Big Data*. Berkeley, CA: Apress, 2018. Disponível em: https://doi.org/10.1007/978-1-4842-3147-0_1. Acesso em: 11 jun. 2020.

SIVARAJAH, U. *et al*. Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, [s. l.], v. 70, p. 263-286, Jan. 2017. Disponível em: https://www.sciencedirect.com/science/article/pii/S014829631630488X. Acesso em: 22 maio 2020.

ZIMMERLE, C. Reactive-based complex event processing: an overview and energy consumption analysis of CEP.js. *In*: SIMPÓSIO BRASILEIRO DE ENGENHARIA DE SOFTWARE, 33., 2019, Salvador. *Proceedings* [...]. Salvador: SBES, 2019. Disponível em: https://dl.acm.org/doi/10.1145/3350768.3352492. Acesso em: 22 maio 2020.

Resumo

Deixar de pensar por nós mesmos e não ver a condição humana nos outros pode ser o primeiro passo que nos leva a uma perigosa apatia que aceita a violência como algo natural, corriqueiro e, até mesmo, justificável. Isso pode esconder uma guerra civil que traz imensos prejuízos à nação. A partir dessa hipótese, utilizando-se das técnicas de mosaico síntese, analisou-se os dados de Agressão – internação (1998-2019) e óbitos (1979-2018), do Brasil, para investigar a suspeita dessa *guerra civil não oficializada* e, a partir dessa hipótese, confrontando-os com os dados de outros 68 países (80,63% da população mundial) e, então, encontrou-se parâmetros para estimar a sua dimensão. Como resultado, identificou-se 1.308.891 vítimas masculinas (94,17%)/81.071 femininas (5,83%). A investigação também mostrou que é possível acabar com uma situação onde, somente em 2015, os Anos Potenciais de Vida Perdidos, chegaram a 2.038.888,00 (homens) e 161.604,50 (mulheres), mas isso é uma mera fração dos 72.339.707,87 Anos Potenciais de Vida Ganhos (APVG), que teríamos, caso conseguíssemos eliminar as mortes por Agressão nos homens. Essa nova medida estatística é obtida a partir do uso das Tábuas de Vida de Múltiplo Decremento e revelou para as mulheres a soma de 7.115.848,59 APVG. Diante desses números, esse trabalho se transforma em um alerta para a necessidade de uma mudança urgente de comportamento e desenvolvimento de uma inteligência emocional na população, para lidar com questões de conflito, na qual as mulheres, provavelmente, detêm a solução. Ter empatia e se importar com o outro é o caminho que apontamos para mudar um quadro de violência que deu ao País um prejuízo, em vidas humanas, de quase US$ 21 trilhões.
Palavras-chave: Big Data. Violência. Endemia.

Abstract

Failing to thinking for ourselves and not seeing the human condition in others can be the first step that leads us to a dangerous threat that accepts violence as something natural, commonplace and even justified. This can hide a civil war that causes immense damage to the nation. Based on this hypothesis, using the techniques of mosaic synthesis, analyzing data on Aggression – hospitalization (1998-2019) and deaths (1979-2018), from Brazil, comparing them with data from 68 other countries (80.63% of the world population) and, then, were found parameters to estimate its dimension. As a result, it was identified 1,308,891 male victims (94.17%) and 81,071 female victims (5.83%). The research has also shown that it is possible to end a situation where, just in 2015, the Potential Years of Life Lost reached 2,038,888.00 (men) and 161.604,50 (women), but this is a mere fraction of the 72,339.707.87 Potential Years of Life Gained (PYLG), that we would have, if we could removed the deaths from Aggression in man. This new statistical measure has been obtained from the use of Multiple Decrement Life Tables and revealed to women a value of 7,115,848.59 PYLG. Given these numbers, this work becomes an alert for an urgent change in behavior and the development of an emotional intelligence in population, to deal with issues of conflict, in which woman is likely to detect a solution. Having empathy and caring for each other is the way to change the situation of violence that has caused the country a loss in human lives of almost US $ 21 trillion.
Keywords: *Big Data. Violence. Endemic.*

Guerra civil?: o que uma análise sobre 40 anos de dados de saúde brasileiros revela

ANDRÉ RENÊ BARONI
Doutor em Epidemiologia, pela Universidade de São Paulo (USP). Professor Pleno da Universidade Estadual de Feira de Santana (UEFS).
barboni@uefs.br

QUANDO HANNAH ARENDT, em 1961, foi enviada pela revista *The New Yorker*, para acompanhar o julgamento do criminoso nazista Otto Adolf Eichmann, capturado em um subúrbio de Buenos Aires na noite de 11 de maio de 1960 por um comando israelense e levado para Jerusalém, ela não tinha ideia de que iria se deparar com a descoberta revolucionária da banalidade do mal (ARENDT, 1999), algo perigoso a que todos nós estamos sujeitos, e não nos damos conta, quando deixamos de pensar por nós mesmos (BARBONI, 2014) e não vemos a condição humana (ARENDT, 2007) nos outros. Nessas circunstâncias, somos capazes de cometer os piores crimes. É como esse tipo de banalização que nos colocou, segundo os dados de morbimortalidade disponíveis no site do DATASUS (BRASIL, 2020), em um estado de verdadeira "guerra civil" e apatia coletiva que este texto trata.

Na época do julgamento de Eichmann, havia um grande burburinho entre a comunidade judaica do mundo inteiro e Hannah Arendt esperava encontrar um "monstro" assassino com fortes convicções antissemíticas, mas, em Jerusalém, ela se deparou com um

O que chama a atenção é que os números brasileiros não são típicos de um país pacífico. Eles chocam não só pelos seus valores, mas pelo longo tempo em que esses indicadores crescentes estão afetando perceptivelmente o nosso perfil populacional

homem comum, um burocrata resfriado que achava simplesmente que se esmerava em cumprir bem a sua função de embarcar aquelas pessoas nos trens. Ele não via que esse trabalho iria conduzir "seres humanos" para a morte certa nos campos de extermínio nazistas. Ele não demonstrava qualquer traço de racismo/antissemitismo. Ele apenas tentava ser eficiente no seu trabalho e isso chamou a atenção dela (ARENDT, 1999).

Hoje em dia, as pessoas vão aos cinemas, assistem a filmes na TV e jogam videogames onde os "heróis" cometem as maiores atrocidades contra os "vilões", essas pessoas não se chocam com isso, elas até ficam felizes e torcem para que esses "heróis" cometam esses atos de violência. Depois... elas simplesmente retomam a rotina das suas vidas diárias sem qualquer remorso ou sentimento de compaixão por aqueles que sofreram nas mãos dos seus "heróis". Por que isso se dá? Hannah Arendt parece ter encontrado a resposta... Um simples mecanismo de desconexão nossa para com o outro que nos torna apáticos com relação ao seu sofrimento?... Mais do que isso, se de alguma forma o outro for visto como algum tipo de encarnação do mal, "está automaticamente justificada toda e qualquer violência contra ele". Cometê-la parece nos tornar verdadeiros "heróis".

A falta de empatia está na base da violência e é o primeiro passo rumo a ela. Esse trabalho suspeita que estamos vivenciando, no Brasil, uma verdadeira guerra civil não oficializada e analisa historicamente os dados de morbimortalidade por Agressão e demais Causas Externas no Brasil, disponíveis no *site* do DATASUS (BRASIL, 2020), comparando-os, criticamente, com os dados de alguns outros países. O que chama a atenção é que os números brasileiros não são típicos de um país pacífico. Eles chocam não só pelos seus valores, mas pelo longo tempo em que esses indicadores crescentes estão afetando perceptivelmente o nosso perfil populacional.

Esse trabalho faz parte de um projeto de democratização da informação em saúde e popularização da ciência. Baseia-se em uma nova forma de pensar a ciência, não mais como o campo de saber por excelência, mas como um dos quatro campos de fragmentação do conhecimento, tão importante quanto os demais: filosofia, religião e arte. Essa forma de pensar a ciência se pauta no "homem integral" que além da razão e da sensação, também valoriza a intuição e o sentimento como elementos importantes para se chegar a um conhecimento mais profundo e verdadeiro.

Sua linha investigativa se baseia nas técnicas de um mosaico síntese, como aqueles utilizados pelos investigadores policiais. As pistas que a investigação vai revelando são dispostas em um grande painel onde se procura traçar correlações que possam revelar a verdade do crime cometido. Usam-se todos os recursos de que se dispõe. Fazem-se su-

posições e investigam-se os caminhos que elas levam. Testam-se a coerência dos fatos com as decorrências dessas suposições e em algum momento dessa análise e síntese investigativa a verdade se revela e o crime se esclarece. Em resumo, isso descreve o método investigativo.

Portanto, nos parece adequado utilizar essa técnica para investigar a questão da violência no Brasil e, mais especificamente, as agressões que resultam em morte, para tentar dimensionar: 1 – o número de vítimas que essa guerra civil não oficial produziu em 40 anos; 2 – os Anos Potenciais de Vida Perdidos (APVP); 3 – os Anos Potenciais de Vida Ganhos (APVG), que teríamos caso conseguíssemos eliminar as mortes por Agressão (obtido através das Tábuas de Vida de Múltiplo Decremento)e, 4 – o valor financeiro que essas mortes representariam em dólares americanos.

METODOLOGIA

Trata-se de um estudo epidemiológico de base populacional e uma reflexão crítica apoiada nos dados de população, óbito (1980-2018) e internações (1998-2019) por Agressão e demais Causas Externas (CID-BR-10), segundo o sexo e faixa etária, a partir dos dados disponíveis no *site* do DATASUS (BRASIL, 2020). Além das técnicas e ferramentas da análise descritiva, foram utilizadas as tábuas de vida de múltiplo decremento (BARBONI, 2002) para estimar o impacto que esses grupos de causa de morte têm na esperança de vida do povo brasileiro.

A base da técnica investigativa utilizada foi o mosaico síntese desenvolvida por Barboni (2014), que conjuga os recursos de análise e síntese até que se consiga, por intuição, encontrar a solução para o problema.

Assim sendo, como pistas iniciais nos valemos das informações disponibilizadas no site do DATASUS (BRASIL, 2020) (séries históricas dos dados de população, internação e óbitos por sexo e faixa etária). Utilizando apenas um computador conectado à internet, obteve-se acesso a esses dados que foram trabalhados na planilha eletrônica LibreOffice Calc, um *software* livre rodado no sistema operacional Linux, distribuição Ubuntu.

Além dos referidos dados, foi utilizada uma planilha eletrônica produzida e disponibilizada pela Organização Mundial de Saúde (WORLD HEALTH ORGANIZATION, 2011)[1]. A partir dos dados de mortalidade por Causa Externa, contidos nessa planilha, elegeu-se 69 países (incluindo o Brasil), que representam 80,63% da população mundial, como pa-

1 www.who.int/entity/gho/mortality_burden_disease/global_burden_disease_death_estimates_sex_age_2008.xls.

Isso nos dá uma ideia do impacto que uma política bem sucedida de combate à violência teria na mudança do perfil demográfico da população brasileira e serve de mais um incentivo para um esforço coletivo de implementá-la

râmetro para analisar os dados do Brasil e saber se vivemos ou não em um estado de guerra civil não oficializada. Caso essas suspeitas se confirmassem, queríamos saber se existe no mundo pelo menos um país onde a mortalidade por violência entre os homens não fosse tão desigual com relação às mulheres e quais seriam os possíveis parâmetros que poderíamos utilizar para se chegar a um perfil de mortalidade para um Brasil em paz.

A investigação da série histórica dos óbitos por Causas Externas, no Brasil, complementa esse trabalho na medida em que podemos cruzar as informações dos óbitos por Agressão (intencionais/não intencionais) com os óbitos por Acidente de Transporte (a princípio, não intencionais).

De posse desses parâmetros, podemos, então, chegar ao novo perfil de mortalidade esperado e possível para um Brasil em tempos de paz e utilizar as técnicas de construção de Tábuas de Vida de Múltiplo Decremento (BARBONI, 2002), para estimar a Esperança de Vida com e sem a eliminação do risco de morrer por essas causas. Isso nos dá uma ideia do impacto que uma política bem sucedida de combate à violência teria na mudança do perfil demográfico da população brasileira e serve de mais um incentivo para um esforço coletivo de implementá-la.

Há um problema, porém: essas técnicas exigem que os dados de população e óbito sejam confiáveis e o sub-registro de óbitos, no Brasil, nos obriga a fazer correções para se evitar a superestimação da esperança de vida na população. Valeu-se do material produzido pelo Instituto de Brasileiro de Geografia e Estatística (2013) para tentar definir esses fatores de correção de óbitos para os dados disponibilizados no site do DATASUS (BRASIL, 2020). Segundo esse trabalho, considera-se que a partir de um ano o sub-registro de óbitos é constante. Construindo, então, a Tábua de Vida para o ano censitário de 2010 com os dados sugeridos de correção (1,06 – homens e 1,10 – mulheres) e ajustando os dados de óbitos informados no site do DATASUS com os óbitos das tabelas construídas pelo Instituto Brasileiro de Geografia e Estatística (2013) percebeu-se que o melhor ajuste se dava com um fator de correção de 1,1544 e 1,17245, respectivamente, para as populações masculina e feminina menores de um ano e 1,068 e 1,0975 para quem tinha um ano ou mais, para homens e mulheres nessa ordem. Usando esses valores, pode-se estimar o sub-registro de óbitos para o ano de 2015. Os óbitos por Causa Externa são considerados sem sub-registro e, portanto, para eles o fator de correção é 1, independentemente da idade ou do sexo.

Além das Tábuas de Vida de Múltiplo Decremento valeu-se de dois indicadores: os Anos Potenciais de Vida Perdidos (APVP) e os Anos

Potenciais de Vida Ganhos (APVG). O primeiro, é clássico e foi utilizado por Andrade e Mello-Jorge (2013) para analisar a mortalidade brasileira por Acidentes de Transporte em 2013. Baseou-se nesse trabalho para estimar os APVP, no Brasil, em 2015, por todas as causas de óbito, pelas Causas Externas, pelos Acidentes de Transporte e pelas Agressões. Utilizou-se a mesma técnica referida nesse trabalho, porém, entendeu-se que arbitrar uma idade (70 anos) como sendo a idade ideal a se atingir, como esse método implementa, é algo que não traduz o que de fato se perde por não se ter implantado uma Política que evitasse esses óbitos.

Os Anos Potenciais de Vida Ganhos (APVG) é a nossa proposta de um indicador positivo que nos dá uma noção melhor do que ganharíamos se tal política fosse implementada. Ele nos dá uma ideia melhor do tamanho do "filão de ouro" que temos a ser explorado, ao trabalhar pela melhoria da qualidade de vida da população brasileira, e ao invés do somatório dos anos perdidos para cada pessoa que morreu (entre 1 ano e 70 anos de idade), preferimos trabalhar com o somatório do tempo de vida ganho (diferença entre a esperança de vida com e sem a eliminação do risco de morte) por cada pessoa, em cada faixa etária da população. Isso parece bem mais estimulante para nos incentivar um comprometimento maior com uma Política de real combate à Violência. Aquele esforço que todo mundo sabe que tem de fazer, mas que implica em mudanças de hábitos que o nosso comodismo/egoísmo sempre deixa para um depois e que raramente chega. Para isso precisamos das Tábuas de Vida de Múltiplo Decremento cuja metodologia pode ser vista na tese de doutorado de Barboni (2002).

Greenstone e Nigam (2020), analisaram o quanto as medidas de distanciamento social para enfrentamento da pandemia de covid-19 nos Estados Unidos da América representaram, em termos de vidas humanas e, arbitrando um valor financeiro para essas vidas, eles puderam estimar os benefícios monetários que essas duras medidas representaram na balança da economia. Com base neste trabalho e dado que foi possível estimar as vítimas desta "guerra civil", por faixa etária, foi possível estimar o montante que essas vidas humanas representariam em termos financeiros. Embora reconheçamos que uma vida humana não tem preço, atribuir um valor monetário a ela pode nos ajudar compreender que o prejuízo das vidas humanas perdidas representa um valor bem maior do que o prejuízo financeiro de uma medida como o isolamento social devido a uma pandemia como a do covid-19.

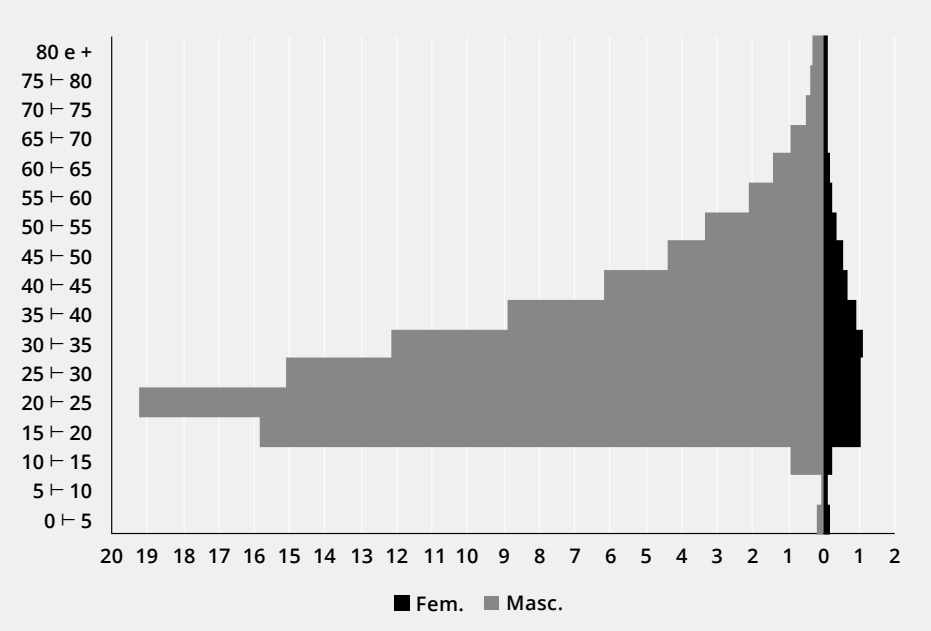
Aplicando isso ao nosso caso, podemos ter mais uma ideia do que a nossa inação/apatia tem gerado e, talvez, assim, para aqueles que só conseguem sentir pelo bolso, se sintam compelidos, finalmente, a fazer algo a respeito.

Quem já ministrou aulas para adultos jovens já deve ter observado como eles tendem a responder questões de prova baseados no senso comum, altamente influenciados pela mídia leiga

RESULTADOS

Quem já ministrou aulas para adultos jovens já deve ter observado como eles tendem a responder questões de prova baseados no senso comum, altamente influenciados pela mídia leiga. Em 2017, diante do gráfico da figura 1, no qual os alunos deveriam se basear para pensar uma política pública de combate à violência no Brasil, um aluno afirmou corretamente que a população masculina é a maior vítima e o maior agressor, mas que desde muito cedo os meninos são mais agressivos que as meninas e, a partir daí, esse aluno se dedicou a descrever uma política de combate à violência contra a mulher. Os homens são mais agressivos! Isso é da natureza deles! Portanto, não percamos tempo com eles e vamos cuidar das mulheres! Será que é isso o que esse aluno pensou? Será isso mesmo um fato, ou toda violência é evitável? O que torna esse gráfico tão assimétrico? Devemos nos conformar com isso e assumir uma política pública que fecha os olhos para 90% das vítimas, porque isso é da "natureza" delas?

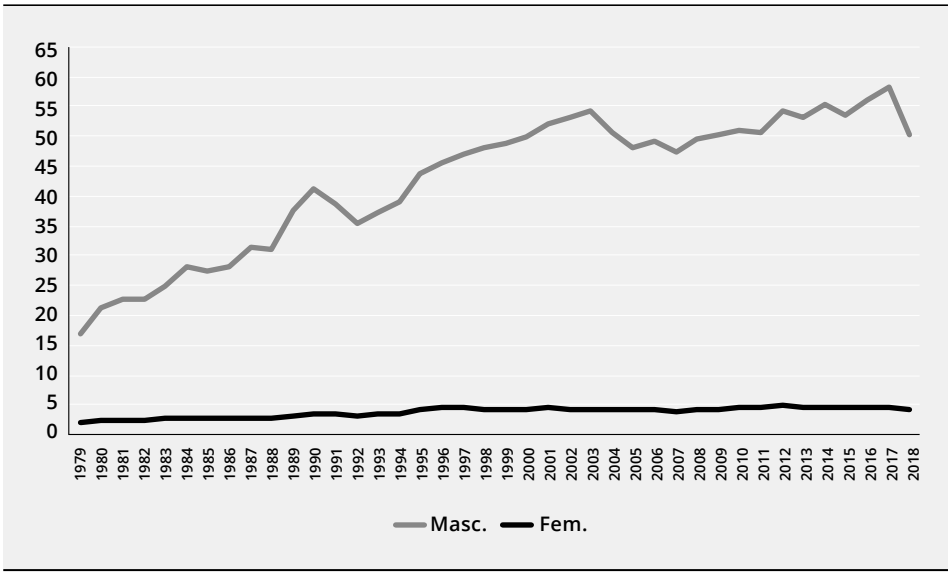
Figura 1
Óbitos por Agressão, segundo o sexo e faixa etária – Brasil – 2015



Fonte: MS/SVS/CGIAE – Sistema de Informações sobre Mortalidade – SIM.

De fato a violência contra a mulher tem crescido e isso é preocupante, mas a violência contra a população masculina, no Brasil, tem indicadores ainda mais alarmantes e, proporcionalmente, ela cresce de forma mais acentuada nos homens do que nas mulheres, pelo menos no que diz respeito à sua forma mais grave, a que leva ao óbito (Figura 2). Será que do ponto de vista da morbidade esse quadro é diferente?

Figura 2
Série histórica dos óbitos por Agressão, para cada 100.000 homens/mulheres – Brasil – 1979-2018

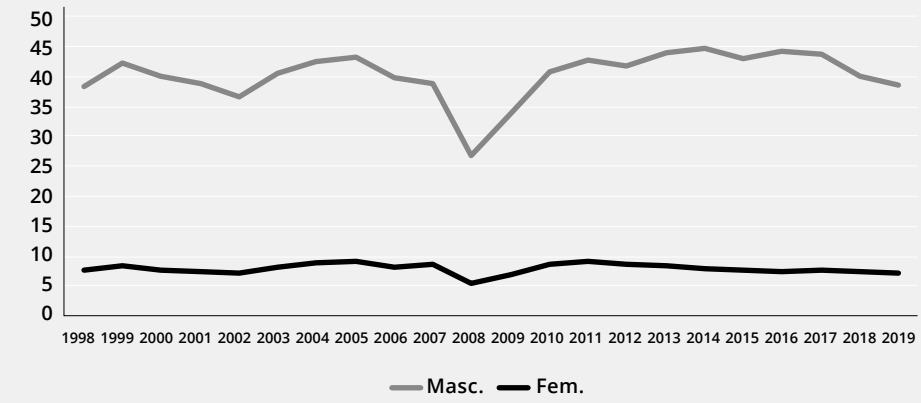


Fonte:MS/SVS/CGIAE – Sistema de Informações sobre Mortalidade – SIM e Censos, Contagem Populacional e Estimativas do IBGE disponibilizadas no site do DATASUS.

As figuras 3 e 4 parecem desmentir isso! Então, a violência contra o homem é também uma questão de saúde pública. Nada anula ou desmerece a importância de todo o trabalho que tem sido feito no combate à violência contra a mulher. Muito pelo contrário, ele deve ser incentivado, mas não se pode tratar esta questão unilateralmente. Em todo agressor existe um problema que precisa ser tratado. Se não cuidarmos das causas, do que leva uma pessoa a agredir outra, o problema em si não será resolvido. As figuras 1 e 4 mostram que, pelo menos, as mulheres parecem ter mais inteligência emocional que os homens. Isso parece ser um bom ponto de partida para começarmos!

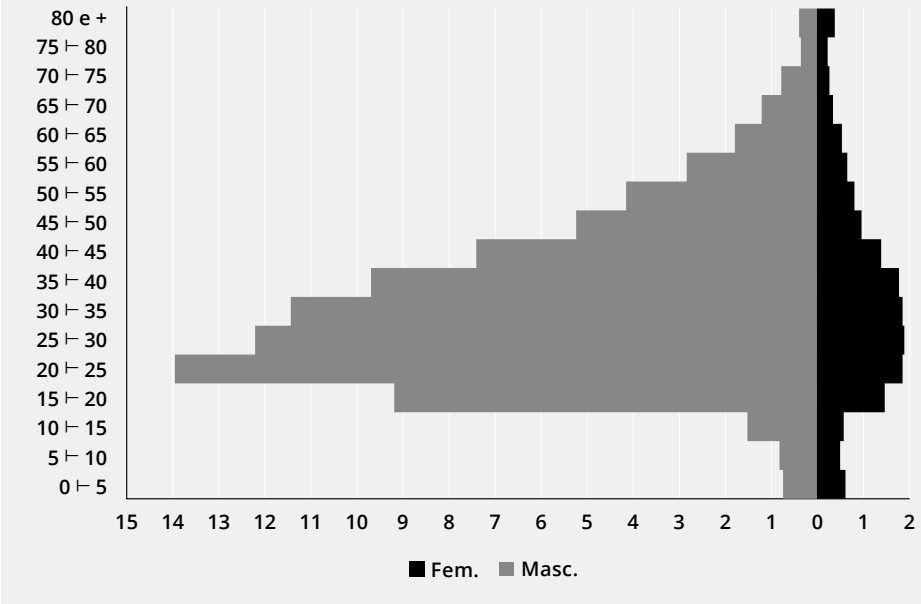
Será que esse perfil de agressividade no qual os homens são as maiores vítimas e os maiores autores da violência se repete no mundo? Será que não existem países onde as figuras 1 e 4 não seriam tão assimétricas? Ao tentar responder essas questões, consultou-se o site da Organização Mundial de Saúde (OMS) e baixou-se uma planilha de mortalidade e carga de doença estimados para os países membros da OMS em 2008 (WORLD HEALTH ORGANIZATION, 2011). Nessa planilha os dados foram agrupados em três faixas etárias de óbito: 1 – de 0 a 14 anos; 2 – de 15 a 59 anos; 3 – 60 anos e mais. A OMS utilizou um código de cores para identificar a qualidade dos dados apresentados: azul – dados razoavelmente completos; azul claro – registros de dados incompletos; rosa – países com informações de causa de óbito não disponível para a maioria das causas. O Brasil foi incluído no grupo azul claro.

Figura 3
Internações no Sistema Único de Saúde por Agressão, para cada 100.000 homens/mulheres – Brasil – 1998-2019



Fonte: Ministério da Saúde - Sistema de Informações Hospitalares do SUS (SIH/SUS).

Figura 4
Internações no Sistema Único de Saúde, por Agressão, segundo o sexo e faixa etária – Brasil – 2019



Fonte: Ministério da Saúde - Sistema de Informações Hospitalares do SUS (SIH/SUS).

Essa investigação concentrou-se, inicialmente, nos casos de "lesões intencionais" e, mais especificamente, nos dados de violência. Priorizou-se os países incluídos no grupo azul e, com isso, conseguiu-se selecionar trinta países para confrontar os dados com o Brasil: Federação Russa; México; Estados Unidos da América; China; Venezuela; Argentina; Chile; Croácia; Coreia do Sul; Cuba; França; Itália; Reino Unido; Canadá; Espanha; Polônia; Alemanha; Israel; Austrália; Bélgica; Bulgária; Grécia;

Holanda; Japão; Portugal; Uruguai; Áustria; Nova Zelândia; Suécia e Suíça. Destes, somente a China pertencia ao grupo azul claro, os demais eram todos do azul.

O Brasil se mostrou como o mais violento deles e os demais foram classificados em ordem decrescente de violência (conforme a ordem de citação no parágrafo anterior). Em um segundo momento, procurou-se incluir na nossa investigação os países que tiveram perdas em operações de guerra[2].

Com isso chegou-se a um total de 69 países para trabalhar (eles representam 80,63% da população mundial e apresentam basicamente a mesma pirâmide etária). Os óbitos por violência e guerra, nos países em guerra, foram inferiores aos homicídios no Brasil. Isso pareceu confirmar as suspeitas do Brasil ser um país em guerra civil não oficializada, mas real, e que a população já se acostumou com ela a ponto de não notá-la e se ver como um povo pacífico e alegre.

Em termos de homicídios (Agressão e/ou Violência), o Brasil aparece em primeiro lugar nesse ranking mundial, com valores típicos de um país em guerra civil das mais acirradas

Em termos de homicídios (Agressão e/ou Violência), o Brasil aparece em primeiro lugar nesse *ranking* mundial, com valores típicos de um país em guerra civil das mais acirradas, mas quando se incluem os dados dos óbitos em operações de guerra e suicídios ele é superado pela Índia e pela China. Somente quando analisamos todas as Causas Externas é que a Federação Russa consegue superar o Brasil.

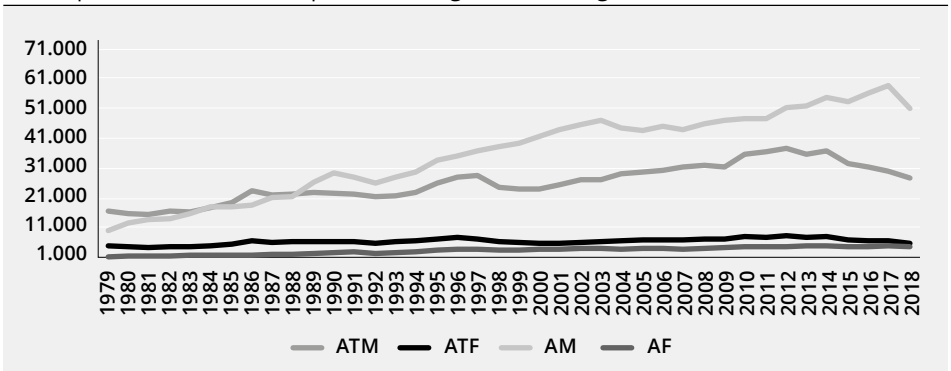
Os dados apontaram para o fato de que a redução significativa dos óbitos por Causas Externas é um grande desafio que os países do BRICS terão que enfrentar. No Brasil, o maior esforço diz respeito à Agressão; Índia, China e Federação Russa terão que lidar com o sério problema do Suicídio. Os Acidentes de Transporte são um grave e crescente problema mundial para todos. Não basta medidas na área da segurança pública! É fundamental atuar nas causas da violência e desenvolver uma cultura de paz entre a população, desde a mais tenra idade. As pessoas precisam cultivar o hábito da gentileza. É imperativo desenvolvemos a nossa "inteligência emocional" para lidarmos melhor com as situações de conflito e as mulheres parecem ter mais experiência acumulada com isso, pois, vivendo sob as mesmas condições socioeconômicas, seus dados são sempre melhores. Os homens precisam aprender com elas.

2 Em ordem decrescente do indicador são eles: Índia; Iraque; Síria; Colômbia; Federação Russa; República Democrática do Congo (Zaire); Filipinas; Etiópia; Estados Unidos da América; Indonésia; Sudão; Nigéria; Paquistão; Afeganistão; Uganda; Quênia; Costa do Marfim; Tailândia; Somália; Myanmar (Birmânia); Angola; Camboja (Kampuchea); Turquia; Argélia; Chade; Nepal; Iêmen; Burundi; Zimbábue; República Centro-Africana; Mali; Haiti; Geórgia; França; Israel; Polônia; Líbano; Estônia; Letônia; Lituânia; Holanda; Portugal; Croácia; Kuwait; Noruega. Estônia, Letônia, Lituânia, Croácia, Kuwait e Noruega também pertencem ao grupo "azul", Índia, Síria, Colômbia, Filipinas, Tailândia, Turquia, Haiti e Geórgia pertencem ao "azul claro", os demais, que não estão no gráfico da figura 5, pertencem ao grupo "rosa".

Imaginem como seria viver no Brasil se conseguíssemos a façanha de fazer com que cada barra à esquerda nos gráficos das figuras 1 e 4 fossem, no máximo, o dobro da barra correspondente à sua direita.

Essa é uma ideia confortadora e nós resolvemos ter uma noção do que essa guerra civil não oficializada, no Brasil, nos tirou nesses 40 anos, caso assumíssemos como naturais os óbitos por Agressão contra o sexo masculino, no máximo, o dobro da Agressão fatal cometida contra o sexo feminino. A figura 5 apresenta a série histórica dos óbitos por Acidentes de Transporte e Agressão, no Brasil, no período de 1979 a 2018. O que chamou a nossa atenção nesse gráfico foi a tendência para os óbitos por Agressão ultrapassarem os óbitos por Acidentes de Transporte, em ambos os sexos.

Figura 5
Óbitos por Acidentes de Transporte (AT) e Agressão (A), segundo o sexo – Brasil – 1979-2018

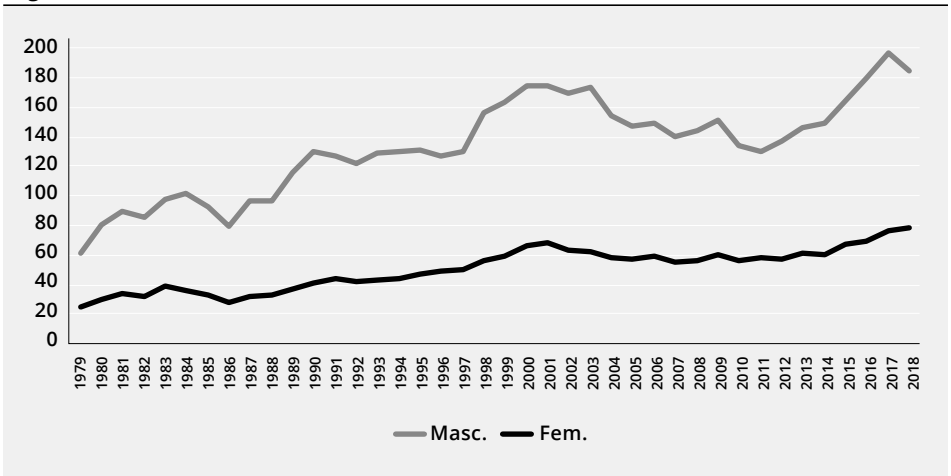


Fonte: MS/SVS/CGIAE – Sistema de Informações sobre Mortalidade – SIM.

Na população masculina, os óbitos por Agressão ultrapassaram os óbitos por Acidentes de Transporte em 1989. Na feminina isso, provavelmente, acontecerá daqui a 20-40 anos, se considerarmos a taxa com que a proporção dos óbitos femininos por Agressão aumenta em relação aos óbitos por Acidente de Transporte (Figura 6).

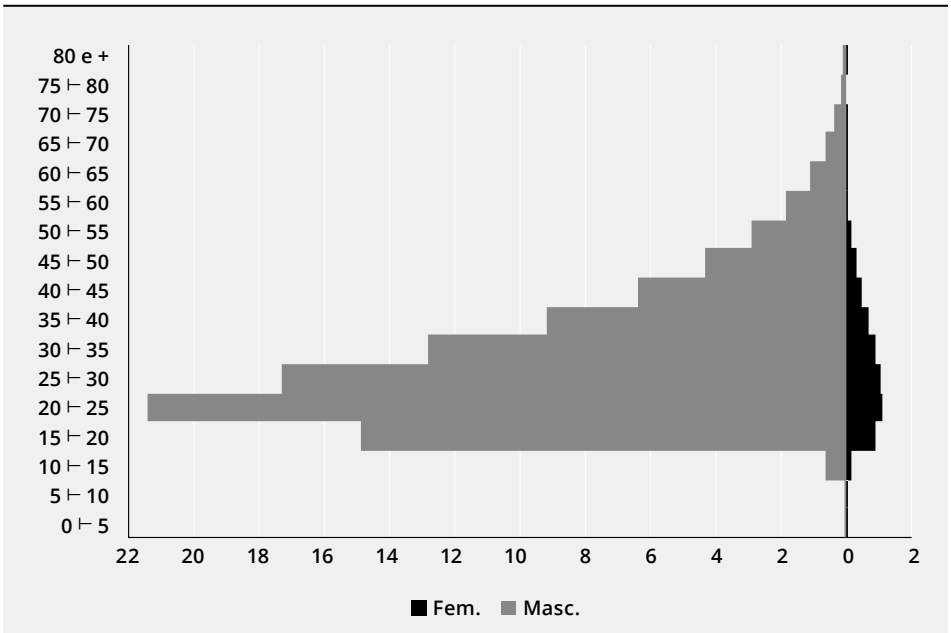
É razoável admitir, então, que não existe intencionalidade nos óbitos por Acidente de Transporte. Eles, no entanto, estão aumentando, isso é fato! Portanto, podemos considerar que os óbitos por Agressão, que seriam "aceitáveis" em um país pacífico, deveriam ser também "não intencionais". Admitimos que é complicado tentar definir uma medida para isso, mas analisando as séries históricas, das figuras 5 e 6, resolveu-se arbitrar um valor aceitável de, no máximo, 20% para a mortalidade feminina por Agressão em relação à mortalidade feminina por Acidentes de Transporte. Esse é um possível valor que se tinha nos anos 1970. Antes das baixas dessa "guerra" começarem a ser computadas pelo Ministério da Saúde! Ou, pelo menos, quando a "guerra" ainda era menos acirrada.

Figura 6
Proporção dos óbitos por Agressão em relação aos óbitos por Acidentes de Transporte, segundo o sexo – Brasil – 1979-2018



Fonte: MS/SVS/CGIAE – Sistema de Informações sobre Mortalidade – SIM.

Figura 7
Distribuição dos óbitos por Agressão, que teriam sido evitados se entre as mulheres, esses óbitos não superassem 20% dos óbitos por Acidentes de Transporte e os óbitos masculinos fossem no máximo o dobro dos femininos, em cada faixa etária – Brasil – 1979-2018



Fonte: MS/SVS/CGIAE – Sistema de Informações sobre Mortalidade – SIM.

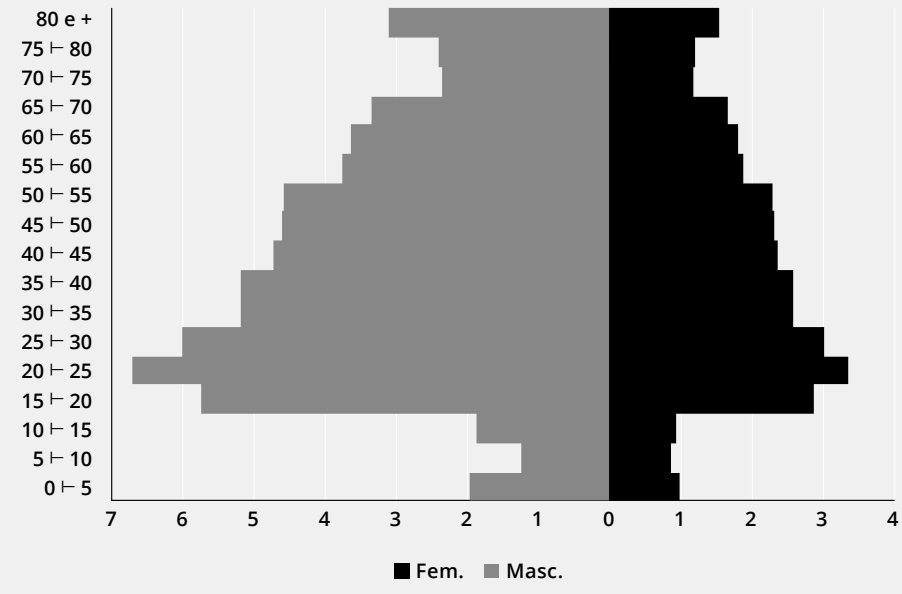
Com base nessa suposição, podemos facilmente calcular os óbitos femininos por Agressão que não deveriam ter existido, para cada faixa etária, e a partir deles, também definimos que os óbitos masculinos, em cada faixa etária, deveriam ser, no máximo, o dobro do feminino. Algo que seria mais condizente com a ideia de uma "violência natural". Isso

Reparem como as diferenças entre gênero se assemelham e como os números ficam próximos quando eliminamos, no Brasil, os óbitos por Agressão que seriam devidos a uma guerra civil não oficializada

resultou, para esse período, num total de 81.071 (5,83%) óbitos femininos e 1.308.891 (94,17%) óbitos masculinos cuja distribuição, por faixa etária, pode ser vista na figura 7.

Voltando à questão formulada naquele semestre letivo, se a violência contra os homens, no Brasil, fosse o dobro da que se comete contra as mulheres, então, o gráfico da figura 1 deveria ser igual ao gráfico da figura 8 (66,5% – masculino e 33,5% – feminino).

Figura 8
Distribuição de óbitos por Agressão, segundo o sexo e faixa etária, descontados os óbitos da nossa suposta "guerra civil não oficializada" – Brasil – 2015



Fonte: MS/SVS/CGIAE – Sistema de Informações sobre Mortalidade – SIM.

Mas será que essa análise é verossímil? A resposta a essa questão pode estar nos dados estatísticos de algum país entre os outros 68 países que elencamos para o nosso estudo e, mais especificamente, entre os países do grupo azul. Em nossas pesquisas, apenas começamos a investigar esses países. O primeiro país que buscamos dados foi os Estados Unidos da América, mas seus indicadores ainda apontam para um elevado índice de violência. No entanto, já no segundo país investigado (Austrália), encontramos a resposta que procurávamos. A tabela 1 compara as taxas de mortalidade por causas externas, por faixa etária e sexo, entre aquele país e o nosso. Reparem como as diferenças entre gênero se assemelham e como os números ficam próximos quando eliminamos, no Brasil, os óbitos por Agressão que seriam devidos a uma guerra civil não oficializada (óbitos por Causas Externas – óbitos por Agressão + óbitos por Agressão Estimados, caso os óbitos femininos

por Agressão não superassem os 20% dos óbitos por Acidentes de Transporte e os óbitos masculinos por Agressão fossem, no máximo, igual ao dobro dos respectivos óbitos femininos por Agressão nessa nova condição).

Se aplicarmos essa técnica aos dados do Brasil, podemos ter, também, uma noção do quanto conseguiremos incrementar a Esperança de Vida em cada faixa etária

Tabela 1
Taxas específicas de óbitos por Causas Externas/100.000, segundo o sexo e faixa etária – Brasil e Austrália – 2012

	Brasil			Austrália		
	Masculino	**Feminino**	**Total**	**Masculino**	**Feminino**	**Total**
0 ⊢ 5	20,27	14,02	17,20	7,10	6,70	6,90
5 ⊢ 15	13,44	6,53	10,04	4,50	2,80	3,60
15 ⊢ 25	82,20	16,19	49,33	39,00	15,80	27,70
25 ⊢ 45	95,19	16,92	55,27	54,70	18,60	36,70
45 ⊢ 65	105,34	23,05	62,17	57,10	22,70	39,70
60 e +	177,87	109,04	138,93	184,20	179,90	181,90
Total	**80,14**	**23,64**	**51,31**	**60,70**	**32,10**	**46,20**

Fonte: Dados do Brasil – DATASUS/SIM-IBGE. Dados da Austrália – AIHW National Mortality Database.
OBS.: Os dados do Brasil consideram os óbitos por Causa Externa ocorridos no ano de 2012, menos os óbitos por Agressões ocorridas no mesmo ano, mais os óbitos por Agressões Estimadas considerando que entre as mulheres eles não poderiam ultrapassar 20% dos respectivos óbitos por Acidentes de Transporte, em cada faixa etária, e entre os homens, os óbitos por Agressões seriam, no máximo, igual ao dobro dos respectivos óbitos femininos em cada faixa etária de cinco em cinco anos. Depois os dados foram agrupados de acordo com as faixas etárias da tabela. Os dados da Austrália se referem ao período 2012-2013.

Até agora, nos valemos apenas de técnicas elementares de epidemiologia e estatística para evidenciar esse estado de guerra civil não oficializada, que o Brasil tem vivido nas últimas décadas, e, também, temos indícios de que é possível se construir uma política para resolver esse sério problema de saúde pública de forma a mudarmos o perfil de mortalidade por Agressão, pelo menos, da figura 1 para a figura 8.

As tábuas de vida de múltiplo decremento nos permitem ir além, explorar a hipótese e calcular os Anos Potenciais de Vida Ganhos (APVG), caso um determinado grupo de causa de óbito deixasse de ocorrer. Essa técnica leva em consideração que uma pessoa que deixe de morrer por essas causas pode ainda vir a óbito pelas demais causas. Isso é necessário! Assim, "os APVG são facilmente obtidos pela diferença entre o total de anos vividos, por uma coorte, a partir da idade 0, obtidos das Tábuas de Vida de Múltiplo Decremento, calculadas com e sem a presença dos óbitos decorrentes do grupo estudado" (BARBONI, 2002, p. 18). Se aplicarmos essa técnica aos dados do Brasil, podemos ter, também, uma noção do quanto conseguiremos incrementar a Esperança de Vida em cada faixa etária. Em um país em guerra, é natural o fenômeno de feminilização do perfil populacional, após o término da guerra, um longo período será necessário para reequilibrar essa "balança". Para os gestores do pós-guerra essa técnica é fundamental.

Preferimos nos valer do APVG por espelhar uma medida positiva que mostra o que podemos "ganhar" se trabalharmos para erradicar/diminuir tal agravo e não o que estamos potencialmente "perdendo"

Tabela 2
Esperança de vida, masculina e feminina, com e sem a eliminação do risco de morrer por grupos de causa de morte (CID-10-BR) – Brasil – 2015

Faixa Etária	Esperança de Vida		Causas Externas		Ac. Transporte		Agressões	
	Masc.	Fem.	Masc.	Fem.	Masc.	Fem.	Masc.	Fem.
<1	71,23	78,05	74,28	78,75	71,99	78,24	72,59	78,18
1 ⊢ 5	71,38	78,12	74,45	78,81	72,15	78,3	72,75	78,25
5 ⊢ 10	67,41	74,15	70,47	74,84	68,18	74,34	68,79	74,28
10 ⊢ 15	62,49	69,23	65,53	69,89	63,25	69,4	63,86	69,35
15 ⊢ 20	57,6	64,31	60,59	64,95	58,35	64,48	58,96	64,43
20 ⊢ 25	53,16	59,46	55,71	60,04	53,83	59,6	54,24	59,56
25 ⊢ 30	48,84	54,62	50,85	55,15	49,39	54,75	49,6	54,7
30 ⊢ 35	44,43	49,8	46,02	50,28	44,89	49,91	44,96	49,87
35 ⊢ 40	39,98	45,03	41,26	45,46	40,35	45,12	40,35	45,08
40 ⊢ 45	35,57	40,33	36,59	40,72	35,87	40,41	35,82	40,36
45 ⊢ 50	31,24	35,71	32,04	36,07	31,47	35,78	31,4	35,74
50 ⊢ 55	27,05	31,22	27,69	31,55	27,24	31,28	27,16	31,23
55 ⊢ 60	23,07	26,88	23,57	27,19	23,21	26,93	23,14	26,89
60 ⊢ 65	19,32	22,72	19,72	23,02	19,42	22,76	19,37	22,73
65 ⊢ 70	15,84	18,79	16,18	19,07	15,92	18,82	15,87	18,8
70 ⊢ 75	12,69	15,14	12,97	15,41	12,74	15,17	12,7	15,15
75 ⊢ 80	9,93	11,86	10,17	12,12	9,97	11,87	9,94	11,86
80 e +	7,65	9,07	7,87	9,34	7,68	9,09	7,66	9,08

Fonte: Cálculos produzidos pelo autor.
OBS.: Para o cálculo das tábuas de mortalidade foi usado um fator de correção para os óbitos naturais (todo óbito com exceção das Causas Externas) de 1,1544 e 1,17245, respectivamente. para os homens e as mulheres menores de um ano e de 1,068 e 1,0975, respectivamente, para a população masculina e feminina com um ano ou mais. Para os óbitos por Causas Externas o fator de correção foi 1.

Tabela 3
Anos Potenciais de Vida Ganhos (APVG) e Perdidos (APVP), em função dos óbitos por Causas Externas, Acidentes de Transporte e Agressões – Brasil – 2015

	APVG			APVP		
	Masc.	Fem.	Total	Masc.	Fem.	Total
CE	184.844.780,32	50.778.460,46	243.776.137,26	3.988.100,50	557.435,00	4.545.535,50
AT	48.552.877,78	11.617.781,66	62.758.281,82	999.504,50	198.784,50	1.198.289,00
A	72.339.707,87	7.115.848,59	83.264.246,59	2.038.888,00	161.604,50	2.200.492,50

Fonte: Cálculos produzidos pelo autor.
OBS.: CE – Causas Externas; AT – Acidentes de Transporte; A – Agressões.

Na literatura científica, tem-se trabalhado mais com os Anos Potenciais de Vida Perdidos (APVP), que é o somatório dos anos que ainda seriam necessários para que cada pessoa falecida de uma dada população atingisse uma idade arbitrária X (LAURENTI *et al.*, 1987). Para sensibilizar os gestores a promover ações que incrementem a expectativa de vida da população, como dissemos, preferimos nos valer do APVG por espelhar uma medida positiva que mostra o que podemos "ganhar" se trabalharmos para erradicar/diminuir tal agravo e não o que estamos potencialmente "perdendo". A tabela 3 apresenta os resultados que encontramos para o Brasil no ano de 2015.

As conclusões do trabalho de Andrade e Mello-Jorge (2013, p.1) afirmam que:

> O impacto da alta taxa de mortalidade é de mais de um milhão de anos potenciais de vida perdidos por acidentes de transporte terrestre, principalmente entre adultos em idade produtiva (mortalidade precoce), em apenas um ano, representando extremo custo social decorrente de uma causa de óbito que poderia ser prevenida.

Os dados de 2015, apesar do pouco tempo passado, corroboram esse estudo, mas, de acordo com a tabela 3, os APVP são uma mera fração dos APVG e os óbitos por Agressão, notadamente entre a população masculina, são o grande vilão dessa história. Algo precisa ser feito com urgência para modificar esse grave problema de saúde pública.

Por fim, uma vez que conseguimos estimar a população morta por essa guerra civil não oficializada e com base no trabalho de Greenstone e Nigam (2020), foi possível estimar os custos em vidas humanas desse genocídio (20.98 trilhões de dólares americanos; 19.76 – homens e 1.22 – mulheres), simplesmente multiplicando o total de vidas perdidas em cada faixa etária pelos valores arbitrados no referido trabalho. Embora alguém possa contestar, resolvemos, para fins desse trabalho, que uma vida de um brasileiro vale a mesma coisa que a vida de um americano, nem mais nem menos.

A guerra mais sangrenta do planeta e essas mortes estão sendo tratadas como simples casos de homicídios e não como vítimas do genocídio fratricida que, de fato, está entranhado na história do povo brasileiro e que não é percebido

DISCUSSÃO

A nossa investigação, utilizando a técnica do mosaico síntese, proposta por Barboni (2014), nos permitiu confirmar as nossas suspeitas de que estamos vivendo em um país onde existe uma verdadeira guerra civil não oficializada. Se fosse oficial, essa guerra estaria sendo combatida com mais veemência por todos aqueles que amam a paz, mas do jeito que está sendo conduzida, as famílias mandam os seus filhos para o *front* de batalha sem nenhum preparo ou treinamento militar. Como resultado, temos um verdadeiro massacre. Segundo as estimativas que pudemos fazer com base nos dados do DATASUS (BRASIL, 2020), de 1979 a 2018, essa *guerra* ceifou a vida de, pelo menos, 1.308.891 homens e 81.071 mulheres, segundo o perfil etário apresentado no gráfico da figura 7. Esta é, provavelmente, a guerra mais sangrenta do planeta e essas mortes estão sendo tratadas como simples casos de homicídios e não como vítimas do genocídio fratricida que, de fato, está entranhado na história do povo brasileiro e que não é percebido.

Essa "guerra" faz suas maiores vítimas entre a população masculina de 10 a 49 anos e após os 50 anos a diferença entre gênero dispara, não porque ela cresce assustadoramente, mas porque não existe praticamente mais vítimas entre as mulheres

A resposta, inesperada, que o nosso aluno nos deu diante da questão que formulamos sobre o gráfico da figura 1 nos fez perceber o quanto estamos acostumados com esta violência, que se tornou endêmica, a ponto de acharmos natural a desigualdade entre o lado esquerdo e o direito da figura 1. Nela, o risco de morte entre a população masculina entre 20 e 25 anos, chega a ser 19,95 vezes maior do que o respectivo risco de morte da população feminina da mesma faixa etária. Essa "guerra" faz suas maiores vítimas entre a população masculina de 10 a 49 anos e após os 50 anos a diferença entre gênero dispara, não porque ela cresce assustadoramente, mas porque não existe praticamente mais vítimas entre as mulheres.

Portanto, a nossa investigação parece confirmar que: 1 – o Brasil tem vivido um estado de guerra civil não oficializada há mais de 40 anos; 2 – Isso tem, evidentemente, um forte impacto socioeconômico e na estrutura etária da população; 3 – a violência se tornou corriqueira e banal a ponto de não ser reconhecida pela população e não mobilizar às autoridades o suficiente para eliminá-la; 4 – essa violência não pode ser considerada natural; 5 – convivendo sob o mesmo teto e sob as mesmas condições socioeconômicas, as mulheres parecem ter, pelo menos, mais inteligência emocional do que os homens para lidar com situações de conflito; 6 – a chave para a eliminação da violência, ou pelo menos a sua redução significativa, está em ações educativas que desenvolvam essa inteligência emocional, também, entre os homens; 7 – para quem se diz amante da paz, os números brasileiros são absurdos e inaceitáveis; é preciso, então, fazer algo a respeito e podemos aprender também com outros povos; 8 – a Austrália, com qualquer outro país, tem também lá os seus problemas com violência, mas seus números, pelo menos, indicam que é exequível uma política pública que reduza significativamente os óbitos por Agressão no Brasil, notadamente entre os homens; 9 – o aluno que simplesmente ignorou o lado esquerdo do gráfico da figura 1, parece não ter se dado conta disso; 10 – Se não quebrarmos essa apatia, alimentada pela mídia leiga e científica, o problema não se resolve.

Os APVP são um bom indicador para medir, de certo modo, o que estamos perdendo com essa "guerra", mas os APVG é um indicador ainda melhor para mostrar o quanto deixamos de ganhar mantendo essa "guerra". O Brasil é rico, sem dúvida, mas o seu povo é sofrido e vive sob condições inaceitáveis, principalmente em um solo com tantas possibilidades de crescimento e desenvolvimento. Essa guerra civil não oficializada é apenas o crime, mas onde podemos encontrar o criminoso?

Se quisermos, verdadeiramente, encontrar a resposta para essa questão temos que parar de olhar para os lados e nos perguntar: qual a natu-

reza do crime? Evidentemente se trata de um crime coletivo, então, o criminoso também é coletivo. Isso já é uma boa pista! Mas, o que motiva esse coletivo a praticar esse crime? A que esse crime está associado? Ao tráfico de drogas? Certamente, uma boa parcela das mortes por homicídio no Brasil e no mundo está associada ao tráfico de drogas ilícitas; mas também, podemos acrescentar as drogas lícitas, como as bebidas alcoólicas, que respondem por uma parcela significativa dessas mortes. Por que as pessoas fazem uso dessas drogas? Por puro hedonismo? Para fugir da realidade? Para anestesiar a sua dor?

Também há outras motivações que levam a esses assassinatos. Cobiça, ambição, inveja, ciúmes e uma série de outros sentimentos egoístas ajudam a compor esse painel de dor e destruição. Pietro Ubaldi, um filósofo italiano que viveu e morreu no Brasil, desde 1932, já apontava para a solução desse problema. Para se combater os males provocados pelo nosso hedonismo egoísta somente com o colaboracionismo altruísta. Esse me parece o melhor remédio para curar esse mal que o nosso coletivo tem vivido há séculos. Em toda agressão existe pelo menos duas vítimas: a que sofre a agressão e a que a comete. A carga maior de dor recai sempre sobre o agressor. É isso o que percebemos quando ampliamos a nossa visão e entendemos os meandros das leis naturais que Ubaldi (1939) nos revela.

Parece um contrassenso, mas se não tratarmos o agressor continuaremos a viver as dores da violência. A natureza do homem é boa! Todo bebê é bom, toda criança só quer ser amada e se relacionar bem com as outras pessoas. Então, o que a torna um "monstro"? As circunstâncias? A nossa indiferença? Que não se importa com o desmantelamento da família? Com o desemprego gerado para garantir um lucro indecente? Com a sedução dos jovens pelos prazeres fáceis que os levam a fugir do trabalho honesto de uma vida reta? Qual a nossa parcela de culpa nessa história? O que fizemos, ou deixamos de fazer, para a construção/manutenção desse quadro de dor e sofrimento? O que podemos fazer para mudar uma realidade tão complexa?

Se você quiser fazer algo a respeito, assista ao documentário: *Quem se importa* de Mara Mourão (2010). Ele lhe dará boas ideias do que fazer para entrar nessa rede altruísta e colaboracionista das pessoas que se importam e querem melhorar a realidade em que vivem. Portanto, sem uma resposta definitiva eu lhes conclamo: s*apere aude* (ouse saber) (BARBONI, 2014) e avante Brasil! Vamos nos unir, homens e mulheres, para eliminar essa "guerra"!

REFERÊNCIAS

ANDRADE, S. S. C. de A.; MELLO-JORGE, M. H. P. de. Mortalidade e anos potenciais de vida perdidos por acidentes de transporte no Brasil, 2013. *Revista de Saúde Pública*, São Paulo, v. 50, n. 59, out. 2016. DOI:10.1590/S1518-8787.2016050006465. Disponível em: https://www.scielo.br/scielo.php?script=sci_abstract&pid=S0034-89102016000100241&lng=pt&nrm=iso. Acesso em: 23 maio 2020.

ARENDT, H. *Eichmann em Jerusalém*: um relato sobre a banalidade do mal. Tradução de José Rubens Siqueira. São Paulo: Companhia das Letras, 1999.

ARENDT, H. *A condição humana*. Tradução de Roberto Raposo,. 10. ed. Rio de Janeiro: Forense Universitária, 2007.

BARBONI, A. R. *O impacto de algumas causas básicas de morte na esperança de vida de residentes em Salvador e São Paulo – 1996. 2002*. Tese (Doutorado em Epidemiologia).- Departamento de Epidemiologia, Faculdade de Saúde Pública – USP, São Paulo: 2002. Disponível em: http://cris.uefs.br/media/pdf/barboni_2002.pdf. Acesso em: 23 maio 2020.

BARBONI, A. R. *Filosofia Brasileira*: um sonho ou uma possibilidade? 2014. Trabalho de Conclusão de Curso (Bacharelado em Filosofia) - Departamento de Ciências Humanas e Filosofia da Universidade Estadual de Feira de Santana, Feira de Santana, 2014. Disponível em: http://cris.uefs.br/media/pdf/barboni_2014.pdf. Acesso em: 23 maio 2020.

BRASIL. Ministério da Saúde. Portal da Saúde. *DATASUS*: informações de saúde (TABNET). Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area= 0204&id=11671&VObj=http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/. Acesso em: 23 maio 2020.

GREENSTONE, M.; NIGAM, V. *Does social distancing matter?* Chicago: University of Chicago, Becker Friedman Institute for Economics, 2020. (Working papel, n. 2020-26). Disponível em: http://dx.doi.org/10.2139/ssrn.3561244. Acesso em: 23 maio 2020.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Tábuas abreviadas de mortalidade por sexo e idade*: Brasil, grandes regiões e unidades da federação: 2010. Rio de Janeiro: IBGE, 2013. (Estudos e pesquisas. Informação demográfica e socioeconômica, 30).

LAURENTI, R. *et al*. *Estatísticas de saúde*. 2. ed. São Paulo: EPU, 1987.

QUEM se importa [documentário]. Direção: Mara Mourão. Produção: Tatiana Battaglia e Mara Mourão. Roteiro: Mara Mourão. Direção de Fotografia: Cristiano Wiggers e Dado Carlin. Animações, artes e gráficos: Camaleão Filmes e Citron-vache. Produção executiva: Maurício Dias e Fernando Dias. Produtor Associado: Gullane Filmes. Narração: Rodrigo Santoro. Mamo Filmes e Grifa Filmes, [s. l.], 2010. 1 filme (93min), sonoro, colorido, 16:9 wide-screen (anamórfico).

UBALDI, P. *A grande síntese*. Tradução de Guillon Ribeiro. Rio de Janeiro: Federação Espírita Brasileira, 1939.

WORLD HEALTH ORGANIZATION. *Mortality and burden of disease estimates for WHO member states in 2008*. [S. l]: WHO, 2011. Disponível em: http://www.who.int/entity/gho/mortality_burden_disease/global_burden_disease_death_estimates_sex_age_2008.xls. Acesso em: 23 maio 2020.

Resumo

O alinhamento entre políticas públicas de saúde e tecnologia contribui para aumentar o fornecimento e a qualidade dos serviços de saúde prestados. As inovações tecnológicas possibilitam o acesso às informações de forma ampla e concatenada com as reais necessidades de saúde da sociedade. Considerando o perfil epidemiológico do Brasil e as limitações do serviço público de saúde, este artigo parte do objetivo de explorar algumas aplicações no uso de Big Data pelos sistemas de saúde globais e, de modo comparativo, analisar a experiência brasileira. Com uso da metodologia exploratória foi feito o levantamento de informações por meio da pesquisa bibliográfica. Verificou-se que o uso de novas abordagens tecnológicas contribui para a redução de custos e a melhoria de cobertura e qualidade na prestação de serviços de saúde, podendo propiciar o aumento da eficiência dos recursos empregados e reduzir as ineficiências de gestão. Contudo, o SUS apresenta desafios estruturais que impedem o avanço do uso de novas tecnologias e há uma desarticulação entre o desenvolvimento tecnológico e produtivo da saúde, que impede a eficiência no setor saúde e, consequentemente, a geração de externalidades positivas para as outras áreas sociais.
Palavras-chave: Big Data. Setor Saúde. Políticas Públicas. Inovação. Tecnologia.

Abstract

The alignment between public health policies and technology contributes to increase the supply and quality of health services provided. Technological innovations enable access to information in a broad and concatenated manner with the real health needs of society. Defining the epidemiological profile of Brazil and the permissions of the public health service, this article is part of the objective of exploring some applications that use large volumes of data from global health systems and, comparatively, analyzing the Brazilian experience. Using of the exploratory methodology, the data was made the collection of information through bibliographic research. It was verified that the use of new technological approaches contributes to cost reduction and improvement of coverage and quality in the provision of health services, which can provide or increase the gains in resources used and management inefficiencies. However, SUS presents structural challenges that impede the advancement of the use of new technologies and there is a disconnect between technological and productive development of health, which prevents efficiency in the health sector and, consequently, the generation of positive externalities for other social areas.
Keywords: *Big data. Health Sector. Public Policies. Innovation. Technology.*

Saúde na era do big data: política e planejamento

JOANA AZEVÊDO FRAGA
Mestre e doutoranda em Economia, pela Universidade Federal da Bahia (UFBA).
joanafr1@gmail.com

INARA ROSA DE AMORIM
Mestre em Economia, pela Universidade Federal de Uberlândia (UFU) e doutoranda em Economia, pela Universidade Federal da Bahia (UFBA). Professora Efetiva da Universidade Estadual de Goiás (UEG).
inaraamorim@gmail.com

IVANESSA THAIANE DO NASCIMENTO CAVALCANTI
Doutora e mestre em Economia, pela Universidade Federal da Bahia (UFBA).
ivanessatnc@gmail.com

SETOR SAÚDE é parte importante do desenvolvimento econômico, sendo este um processo de mudança social pelo qual o crescente número de necessidades humanas é satisfeito, através de uma diferenciação no sistema produtivo, gerada pela introdução de inovações tecnológicas (FURTADO, 1964). A saúde é um direito garantido na Constituição Federal de 1988 e elemento estruturante do Estado de Bem-Estar, integrando o sistema de proteção social. Sua dimensão econômica envolve a base produtiva de bens e serviços, capacidade de geração de empregos e mobilização de investimento em inovação/P&D (GADELHA *et al*, 2011; GADELHA, 2012).

O Sistema Único de Saúde (SUS) do Brasil é uma ampla rede que engloba ações e serviços de saúde, cujo reflexo pode ser visto na atenção básica, de média e alta complexidades; serviços de urgência e emergência; na atenção hospitalar; em ações e serviços das vigilâncias epidemiológica, sanitária e ambiental e assistência farmacêutica. Esse sistema parte dos princípios da universalidade, integralidade e acesso igualitário aos serviços de saúde, para que seja assegurada a segurança social (FIOCRUZ, 2018).

A inserção do Big Data nos sistemas de saúde podem gerar ganhos organizacionais e de gestão, impacto nos custos e redução dos problemas existentes

O SUS prevê uma estrutura híbrida de gestão, com funcionamento simultâneo de uma rede pública e gratuita, e uma rede privada complementar, regida por diretrizes do SUS. Essa rede privada se organizou com estabelecimentos de atendimento particulares e com cobertura de planos e seguros de saúde (ARAÚJO, 2014). Dados oficiais demonstram que 71,1% da população brasileira é atendida exclusivamente pelo sistema público, cerca de 150 milhões de indivíduos, enquanto apenas 27,9% tem cobertura por plano de saúde complementar. As regiões Sudeste, Sul e Centro-Oeste possuem as maiores proporções de cobertura, 36,9%, 32,8% e 30,4%, respectivamente, enquanto as regiões Norte e Nordeste possuem 13,3% e 15,5%, respectivamente (PESQUISA NACIONAL DE SAÚDE, 2015).

A cobertura e a qualidade dos serviços de saúde, juntamente com a baixa resolutividade dos problemas de saúde, são um agravo ao bom funcionamento do SUS. A partir do alto grau de complexidade, podem ser pontuados os principais problemas: a) resolução de questões jurídicas quanto ao acesso igualitário e ao atendimento integral; b) aperfeiçoamento da governabilidade do sistema; c) melhoria na articulação entre as redes pública e privada, para evitar duplicações e direcionar os recursos públicos para a população sem acesso aos planos privados; d) melhoria do acesso e da qualidade dos serviços do SUS, ampliando sua cobertura, certificando as instituições de atendimento e avançando na qualificação dos recursos humanos; e) monitoramento e avaliação dos resultados em saúde por parte de instituições avaliadoras externas, cujos conceitos tenham implicações para a política de distribuição de recursos públicos e, f) ampliação do financiamento setorial (BACHA *et al*, 2011).

Diante deste cenário, a coleta e a análise de dados de boa qualidade são essenciais para melhorias na eficácia e eficiência da prestação de serviços de saúde. O recente desenvolvimento de novos métodos, a chamada abordagem Big Data, tem o potencial de permitir o uso de grande base de dados, possibilitando as análises estatísticas de alta potência, contribuindo para o desenvolvimento da saúde pública. Armazenar enormes repositórios de dados e compilar informações numerosas de várias fontes permite descobrir padrões e associações que, de outra forma, seriam impossíveis. Desse modo, a inserção do Big Data nos sistemas de saúde podem gerar ganhos organizacionais e de gestão, impacto nos custos e redução dos problemas existentes.

Neste artigo, o objetivo é explorar algumas aplicações do uso de Big Data pelos sistemas de saúde globais e, de modo comparativo, analisar a experiência brasileira. O debate considerou o perfil epidemiológico do Brasil e os entraves na oferta do serviço público de saúde, fatores estes essenciais para o fornecimento de diretrizes para o uso da nova

tecnologia. Ademais, foram apresentados os desafios estruturais que impedem o avanço do uso da nova tecnologia pelo SUS.

Além desta introdução, mais quatro tópicos compõem este artigo. Os serviços de saúde no Brasil, o perfil epidemiológico e a distribuição da oferta em saúde são apresentados no próximo tópico. Na sequência, o tópico 3 mostra como o Big Data pode ser aplicado no setor saúde, com exemplos em escala internacional e nacional. O quarto tópico apresenta os desafios para o avanço do uso do Big Data no SUS. E, por último, as conclusões finalizam esta discussão.

A tendência da situação de saúde na região das Américas apontou que os diferenciais econômicos entre os países foram decisivos nas variações dos indicadores básicos de saúde e de desenvolvimento humano

SERVIÇOS DE SAÚDE NO BRASIL

Desenvolver ações em saúde pública requer um conhecimento acerca das condições de vida e de trabalho dos indivíduos que integram a sociedade, dos fatores determinantes e condicionantes do processo saúde-doença e suas implicações. O Brasil, como em outros países da América Latina, apresenta um ambiente socioeconômico marcado pela desigual distribuição de renda, elevado grau de analfabetismo e baixo grau de escolaridade e condições precárias de habitação e ambiente (CARVALHO *et al*, 2017).

Em um estudo divulgado pela OPAS (1998, *apud* CARVALHO *et al*, 2017), a tendência da situação de saúde na região das Américas apontou que os diferenciais econômicos entre os países foram decisivos nas variações dos indicadores básicos de saúde e de desenvolvimento humano. Esse estudo mostrou que o Produto Interno Bruto dos países foi determinante para a redução da mortalidade infantil e elevação da esperança de vida, no acesso à água e ao saneamento básico, no gasto em saúde, na fecundidade global e no incremento da alfabetização de adultos.

O ambiente em que o indivíduo está inserido, juntamente com o conjunto de fatores que o cerca, delimitam as condições do estado de saúde. Portanto, se faz interessante analisar a situação dos demandantes de saúde para identificar os problemas de saúde que, mesmo com o desenvolvimento do SUS, não foram eliminados. Concomitantemente, se faz necessário averiguar o setor de serviços, que compreende o desenvolvimento (ciência e tecnologia) e a produção de produtos (aparelhos, instrumentos médicos, medicamentos e vacinas etc.) voltados a atender as necessidades do setor de serviços de saúde. Sendo assim, os próximos subtópicos apresentarão o perfil epidemiológico populacional das famílias brasileiras e como foi realizada a distribuição da oferta de serviços de saúde entre as instituições interligadas ao SUS.

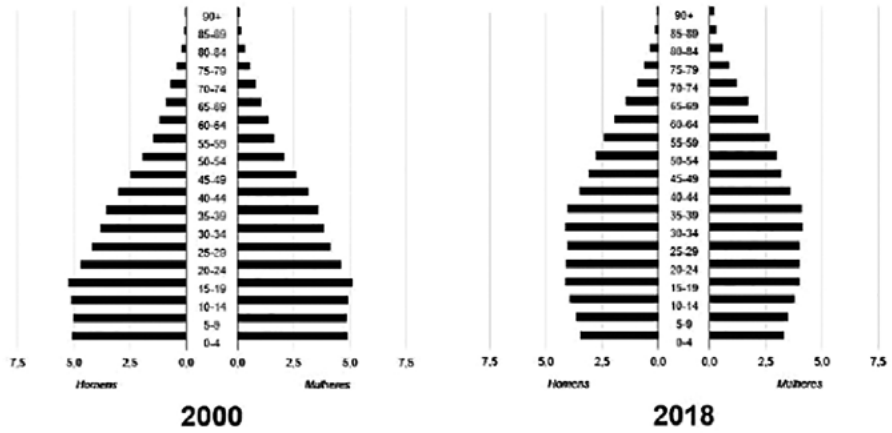
Perfil epidemiológico[1]

Com o envelhecimento da população brasileira, novas problemáticas surgiram, como o aumento de pessoas com doenças crônicas. Além disso, foi possível analisar alguns hábitos culturais da população que inferem na qualidade de vida da mesma. Por meio de dados disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE, 2020), foi possível conhecer os hábitos da população brasileira e inferir como esse processo de envelhecimento influenciará as dinâmicas populacionais.

Segundo a instituição, a estimativa da população nacional até junho de 2020 é de 211,6 milhões de habitantes, dos quais 51,1% são mulheres. A expectativa média de vida é de 76 anos, sendo 80 anos para as mulheres e 73 anos para os homens; 20% da população possuem até 14 anos, 69% tem idades entre 15 e 64 anos e 9% apresentam 65 anos ou mais. A taxa de fecundidade é de 1,7 filho por mulher e a taxa bruta de mortalidade é de 6,5 mortes por mil habitantes.

A projeção para daqui a 26 anos (entre 2017 e 2043) mostra que a população irá atingir o seu limite máximo de 228,4 milhões e passará a decrescer nos anos subsequentes. Conforme as pirâmides populacionais expostas na Figura 1, vemos que há uma tendência de envelhecimento populacional.

Figura 1
Pirâmides etárias – Brasil – 2000/2018



Fonte: Instituto Brasileiro de Geografia e Estatística (2020).

1 Epidemiologia é a designação do estudo das ocorrências dos fenômenos de saúde-doença e seus fatores condicionantes e determinantes nas populações. A epidemiologia permite avaliar a eficácia das intervenções realizadas por meio de saúde pública. Portanto, o perfil ou padrão epidemiológico representa a situação epidemiológica de determinada população.

A pirâmide etária mostra os agrupamentos em barras horizontais de grupos de idades separados por sexo; a parte inferior da pirâmide representa os grupos de faixa etária menor, ao passo que as partes superiores representam a população com idade mais elevada. Atualmente, o Brasil é caracterizado como um país adulto em fase de transição para se tornar um país idoso nos próximos 30 anos. As mudanças identificadas no perfil populacional brasileiro devem-se à diminuição da natalidade ao longo do tempo, à redução das taxas de mortalidade e ao aumento da expectativa de vida.

As mudanças identificadas no perfil populacional brasileiro devem-se à diminuição da natalidade ao longo do tempo, à redução das taxas de mortalidade e ao aumento da expectativa de vida

Com o envelhecimento dos indivíduos, novas doenças surgem e há o agravamento das doenças crônicas. As doenças crônicas representam 72% do total das mortes no País, respondem pelas maiores taxas de morbimortalidade[2] e por cerca de 70% dos gastos assistenciais com saúde. Podemos separá-las em: doenças do coração: 339.066 casos; câncer: 168.562; doenças respiratórias: 59.721; diabetes: 51.828 e outras doenças crônicas: 143.602. Além das doenças crônicas, os hábitos culturais interferem na qualidade de vida das pessoas. Entre os hábitos ruins, estão, 15% das pessoas são fumantes; 34% consomem carnes gordurosas; 18% bebem álcool em excesso; 63% estão acima do peso ou são obesos (BRASIL, 2018b).

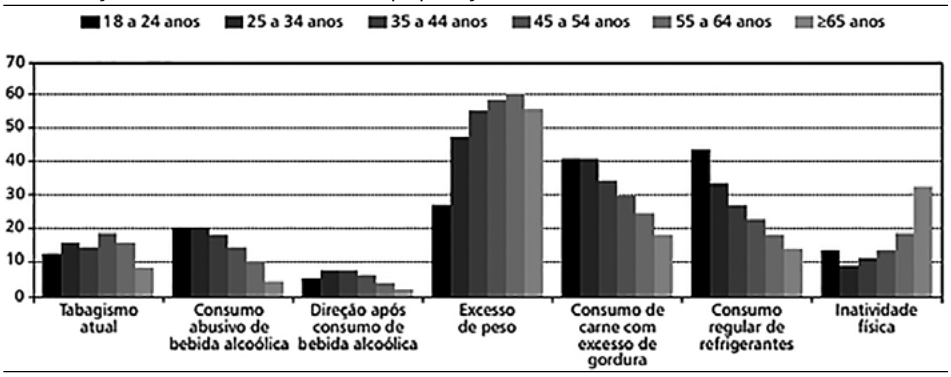
Em uma pesquisa feita pelo Sistema de Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico (Vigitel), implantado pelo Ministério da Saúde no ano de 2011, os entrevistados apresentaram as seguintes estatísticas: mais de 10% da população adulta foi exposta à fumaça do cigarro; 50% estavam acima do peso, destes 15,8% eram obesos; apenas 20% dos entrevistados afirmaram ingerir cinco ou mais porções diárias de frutas e hortaliças; apenas um terço dos adultos pratica 150 minutos de atividade física e 14% foram considerados inativos; 26% dos homens ingeriram de cinco doses ou mais de bebidas alcoólicas em uma mesma ocasião e, destes, 11,4% afirmaram ter dirigido após o consumo de bebidas alcoólicas.

O objetivo da pesquisa foi mostrar que existe uma tendência dos fatores de risco para o desenvolvimento de doenças crônicas aumentar proporcionalmente à faixa etária da população. A pesquisa também identificou que o consumo abusivo de álcool, carnes gordurosas e refrigerantes tende a reduzir com o aumento da idade, como exposto na Figura 2.

2 Morbimortalidade se refere ao índice de pessoas mortas em decorrência de uma doença específica em determinado grupo populacional. Em síntese, é a junção de dois conceitos mobilidade (número de indivíduos portadores de determinada doença em relação ao total da população analisada) e mortalidade (quantidade de pessoas mortas em um grupo específico).

O Brasil tem experimentado um padrão epidemiológico diferente [...] moléstias antigas, como tuberculose, malária e hanseníase, convivem com males da saúde moderna: doenças crônicas e mortes violentas

Figura 2
Distribuição dos fatores de risco na população adulta de acordo com a faixa etária



Fonte: VIGITEL, Brasil (2013).

Outrossim, o Ministério da Saúde (BRASIL, 2013) acredita que é importante conhecer e monitorar a distribuição dos fatores de risco e de proteção à saúde nos adultos, para poder atuar sobre o processo saúde-doença, com o desenvolvimento de políticas públicas direcionadas para melhorar a qualidade de vida da população brasileira. A alternativa para reduzir a carga de doenças no País passa por reduzir os fatores de risco e aumentar os fatores protetores. Essa visão defendida pelo Ministério da Saúde é a mesma verificada em organismos e instituições internacionais.

As mulheres representam os maiores consumidores de assistência à saúde. À medida que a idade avança, a diferença entre a busca por tratamentos médicos entre homens e mulheres tende a se acentuar mais. Há diferenças relacionadas ao nível de instrução: quanto maior a escolaridade, maiores são os cuidados de saúde. A localização geográfica é outro fator que também influencia na busca por serviços de saúde. Nos locais onde o acesso ao serviço de saúde é mais fácil, a população tende a utilizar os serviços. Por exemplo: a população urbana consome mais serviços de saúde do que a população rural (ZUCCHI *et al*, 2000).

Diferentemente do que aconteceu em alguns países latino-americanos, como Chile, Cuba e Costa Rica, e em alguns países desenvolvidos, o Brasil tem experimentado um padrão epidemiológico diferente. A transição epidemiológica está mais para uma superposição de contextos epidemiológicos apresentados ao longo do tempo, uma vez que problemas velhos e novos estão coexistindo de forma desigual nas diferentes regiões do País (CARVALHO *et al*, 2017). Neste sentido, moléstias antigas, como tuberculose, malária e hanseníase, convivem com males da saúde moderna: doenças crônicas e mortes violentas.

A transição do perfil epidemiológico trouxe inúmeros desafios ao sistema de saúde. A alta carga de doenças crônicas e a consequente perda

de saúde da população têm aumentado além da capacidade de cobertura no SUS. Assim, o envelhecimento da população e a transição contrastante do perfil epidemiológico acirram em um novo tipo de demanda por serviços médicos e sociais. Portanto, esse cenário permite diagnosticar um problema central de saúde pelo lado das famílias, a insuficiência da capacidade instalada no SUS. E deste, ramificam diversos outros problemas.

A cobertura não deveria ser vista como apenas uma área geográfica que apresenta uma unidade de saúde, mas sim estar pautada no fornecimento de uma gama de serviços de saúde que precisam ser regulados

Distribuição da oferta em saúde

A oferta de serviços de saúde compreende uma gama de prestadores de serviços que atuam no fornecimento de bens e produtos direcionados à saúde. Os prestadores de serviços incluem as unidades básicas de saúde, hospitais, laboratórios e clínicas de medicina diagnóstica, profissionais da saúde (médicos), fornecedores de medicamentos, fornecedores e distribuidores de equipamentos e materiais médicos.

A cobertura e qualidade dos serviços de saúde, juntamente com a baixa resolutividade dos problemas de saúde, são um grande agravo ao bom funcionamento do SUS. Para o autor, os problemas de cobertura referem-se a: a) ineficiência dos programas de promoção e prevenção de saúde, b) coberturas desiguais e/ou incompletas, c) acesso a políticas de promoção insuficientes, d) ausência de prevenção e serviços, e) falta de equidade no acesso aos serviços, f) baixa qualidade e pouca resolutividade dos serviços e, g) insegurança dos pacientes (MEDICI, 2011).

A cobertura não deveria ser vista como apenas uma área geográfica que apresenta uma unidade de saúde, mas sim estar pautada no fornecimento de uma gama de serviços de saúde que precisam ser regulados. As unidades de saúde devem apresentar medicamentos, pessoal qualificado e equipamento para, no mínimo, prover atendimento ambulatorial. Importante indicativo pode ser observado segundo os estudos do *Projeto Avaliação do Desempenho do Sistema de Saúde* (MONITORAMENTO..., 2019). Entre 2009 e 2017, o número de leitos hospitalares (clínicos, cirúrgicos, pediátricos e obstétricos) sofreu redução: de 1,87 por cada mil habitantes para 1,72 por mil habitantes, número inferior ao estabelecido pela Portaria GM/MS n° 1101/2002 (vigente até 1º outubro de 2015), que era de 2,5 a 3,0 leitos por cada mil habitantes.

Outra questão se refere à qualidade dos serviços. O Brasil apresenta poucos estabelecimentos de saúde com níveis de acreditação aceitáveis. A acreditação é uma avaliação realizada por instituições especializadas que garantem a qualidade da infraestrutura, recursos humanos e gestão condizente com a missão institucional. Entre as empresas que fazem acreditação no Brasil estão a Organização Na-

O crescimento da atenção básica e da rede de urgência e emergência, após o ano 2000, está relacionado ao incremento em UBS [unidades básicas de saúde] e clínicas

cional de Acreditação (ONA) e o Consórcio Brasileiro de Acreditação (CBA). Em 2010, apenas 5% dos estabelecimentos de saúde estavam acreditados, portanto, se faz necessário melhorar a qualidade das instituições de saúde para melhorar a qualidade dos serviços ofertados (MEDICI, 2011).

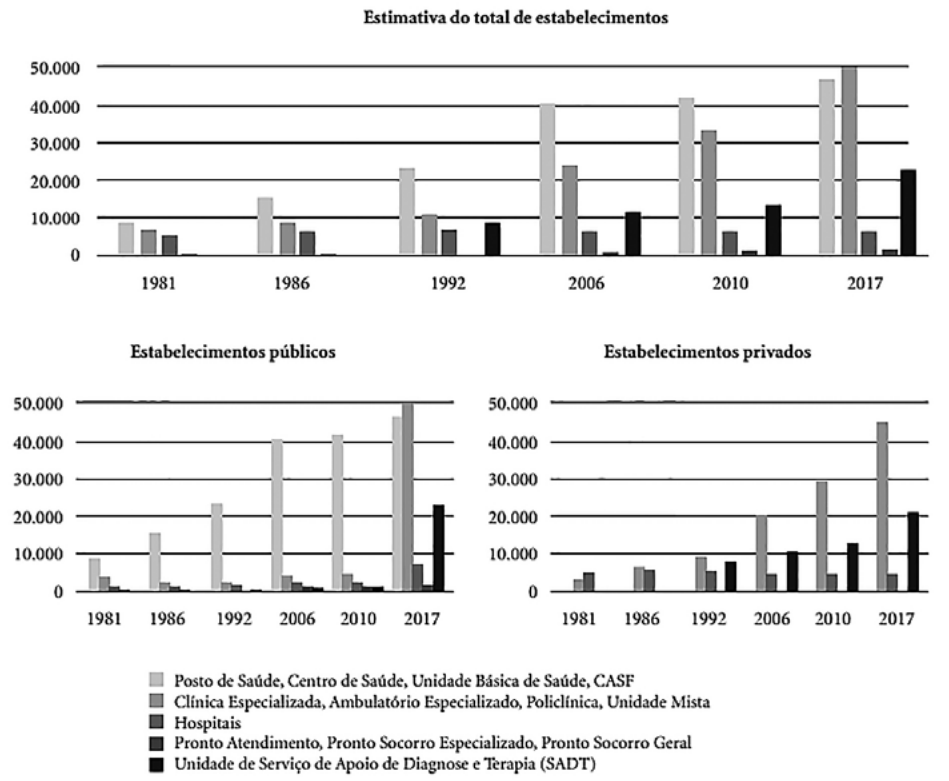
E os problemas de resolutividade dizem respeito tanto à escassez de dados quanto à resolução das questões de saúde que impossibilitam atestar se os serviços ofertados estão contribuindo para a melhoria da vida das pessoas. O imprescindível seria gerar dados que expressem se os problemas de saúde por indivíduo que utilizou o SUS foram resolvidos ou não, para então saber se o sistema está indo no caminho de seus objetivos ou se precisa ser ajustado (MEDICI, 2011).

O número de estabelecimentos de saúde subiu de 21.532 para 129.544, de 1981 a 2017. Nos primeiros anos da análise, houve crescimento das unidades básicas de saúde (UBS) e clínicas, o segmento hospitalar teve crescimento com pouca variação (5.660 para 6.794). Após a década de 1990 os prontos-socorros e as unidades de serviço de apoio de diagnose e terapia (SADT) apresentaram aumento no número das instituições. O crescimento da atenção básica e da rede de urgência e emergência, após o ano 2000, está relacionado ao incremento em UBS e clínicas. Contudo, no ano de 2017, as UBS são, em sua maior parte, públicas (99,2%), enquanto as clínicas são, em maioria, privadas (86,8%). Os estabelecimentos privados também se concentram em hospitais e em SADT (VIACAVA *et al*, 2018).

O setor privado sempre esteve presente na prestação dos serviços de atenção à saúde e durante o processo de reforma sanitária brasileira a partir da década de 1970. A atuação e expansão do SUS estava interligada ao setor privado, principalmente nas relações de serviços conveniados e contratados. Na divisão entre estabelecimentos públicos e privados, apontada na Figura 3, a participação das instituições privadas apresenta-se concentrada em hospitais, pronto atendimento, pronto-socorro especializado, pronto-socorro geral e SADT, o que ressalta a especificidade do segmento. Há certa interdependência entre os setores público e privado na atenção à saúde. O SUS, para garantir atendimento à população, recorre às instituições privadas. Contudo, uma boa parte das instituições privadas depende dos recursos públicos, apresentando uma categoria de uso misto (VIACAVA *et al*, 2018).

A Figura 3 mostra a relação geral e específica por tipo de estabelecimento e a distribuição entre instituições públicas e privadas entre 1981 até 2017. É nítido o aumento da quantidade de instituições de saúde nas diversas áreas, mas principalmente no atendimento clínico.

Figura 3
Evolução da rede assistencial segundo o tipo de estabelecimento e natureza administrativa



Fonte: Viacava e outros (2018).

O tomador de decisão pode possuir mais informações que as demais partes, que também serão afetadas pela decisão

De todo exposto, a oferta nem sempre corresponde às necessidades. Algumas vezes, aparece em função da demanda e em outras pode ser criada uma demanda em função de um determinado bem ou serviço. Essa última forma citada aproxima a saúde a uma estrutura de mercado tradicional, que ressalta algumas características do consumo desnecessário de bens, induzido por quem deveria cuidar das necessidades. Nas relações entre prestadores de serviços e usuários pode haver comportamentos de risco (MALIK, 2001).

O comportamento de risco aparece quando o mercado é incapaz de regular coerentemente. Por exemplo, as diversas falhas de mercado impedem que o mecanismo de preço funcione de forma eficiente. Essas falhas são oriundas, principalmente, da diversidade de informações que os tomadores de decisão possuem. Isto é, o tomador de decisão pode possuir mais informações que as demais partes, que também serão afetadas pela decisão. Esse problema é conhecido como assimetria de informação. O resultado das falhas de mercado pode gerar ineficiência no atendimento de saúde, aumento de preços, apropriação de valores, limitações no acesso aos serviços e perda de qualidade no serviço ofertado (AZEVEDO *et al*, 2016).

A tecnologia envolve uma mudança de bancos de dados relacionais e estruturados que permitem que sejam estabelecidas relações entre dados armazenados em tabelas, para bancos de dados abertos, que podem ser mantidos em diferentes tipos de formatos

Portanto, o novo padrão da população apresenta um duplo desafio, coloca mais pressão financeira sobre o SUS, visto que o sistema enfrenta resistência em mobilizar recursos adicionais sob o atual padrão de financiamento. E pressiona a reorganização da prestação de cuidados de saúde para lidar de forma mais eficaz com as antigas e novas doenças. Na seção subsequente, foi explanado como a utilização do Big Data pode auxiliar na melhoria do sistema público em saúde.

BIG DATA E APLICAÇÕES NA SAÚDE

O contexto de envelhecimento populacional e os diferentes perfis de saúde e problemas da população convivem com a crescente incorporação de tecnologias nos sistemas de saúde, apontando um campo com desafios a serem debatidos. Isso direciona para a necessidade de adequações na estrutura de atenção à saúde às novas demandas e às inovações da área, de forma que seja mantido um sistema de saúde democrático, sendo relevante compreender como ocorrem os processos de mudança (COSTA, 2016).

O Big Data consiste em conjuntos de dados com tamanhos[3] além da capacidade das ferramentas de software comumente usadas para captura, armazenagem, gerenciamento e processamento de dados. Esses dados são oriundos do rápido crescimento do volume e da diversidade de dados digitais gerados em tempo real, como resultado do papel cada vez mais importante desempenhado pelas tecnologias da informação nas atividades cotidianas. O que é definido como grande varia com o tempo, à medida que a tecnologia avança, e pelo setor econômico, dependendo da tecnologia disponível (MCKINSEY, 2011; CEPAL, 2014).

A tecnologia envolve uma mudança de bancos de dados relacionais e estruturados que permitem que sejam estabelecidas relações entre dados armazenados em tabelas, para bancos de dados abertos, que podem ser mantidos em diferentes tipos de formatos, quais sejam: i) estruturados, onde os dados são organizados em sistemas transacionais, com uma estrutura fixa bem definida, e armazenados em bancos de dados relacionais, ii) semiestruturados, onde os dados não têm estrutura regular ou podem evoluir imprevisivelmente, são heterogêneos e podem ser incompletos e iii) não estruturados, onde os dados não foram incorporados às estruturas regulares. Sua manipulação correta permite geração de informações e conhecimento baseados em informações completas em curto período de tempo (CEPAL, 2014).

3 O tamanho das bases de dados manipuladas por Big Data chega à escala dos terabites (10^{12} bytes), petabytes (10^{15}) e zettabytse (10^{21}).

A utilização do Big Data envolve o avanço de tecnologias de informação e comunicação que abrangem banda larga, a internet das coisas, computação em nuvem, gestão de dados, mídias sociais, entre outros, e não apenas uma tecnologia específica (HADOOP, 2014). O Big Data pode ser definido por cinco dimensões, quais sejam: i) volume, que é a quantidade de dados; ii) velocidade na qual os dados são gerados, capturados e compartilhados; iii) variedade de número e natureza dos diferentes tipos de dados; iv) veracidade dos dados, ou seja, sua precisão preditiva, e v) valor, o que implica uma redução em sua complexidade para torná-los utilizáveis na tomada de decisões (COMISIÓN ECONÓMICA PARA AMÉRICA LATINA Y EL CARIBE, 2014). Dessa maneira, pensa-se nesta nova abordagem a partir de um alto potencial de geração de eficiência e efetividade para os sistemas de saúde globais.

A Assembleia aprovou por unanimidade uma resolução abordando a priorização de medidas que promovam o desenvolvimento, a avaliação, a implementação e expansão da utilização das tecnologias digitais como uma forma de promover acesso universal, equitativo e acessível à saúde

Experiências internacionais

Conhecida como e-Saúde ou Saúde 4.0, a saúde digital é uma área que vem ganhando espaço nos debates sobre o futuro dos sistemas de saúde nos últimos anos, principalmente a partir 72ª Assembleia Mundial da Saúde (WORLD HEALTH ORGANIZATION, 2019), organizada pela Organização Mundial de Saúde (OMS). A Assembleia aprovou por unanimidade uma resolução abordando a priorização de medidas que promovam o desenvolvimento, a avaliação, a implementação e expansão da utilização das tecnologias digitais como uma forma de promover acesso universal, equitativo e acessível à saúde. Segundo a OMS, prontuários eletrônicos, bases de dados clínicos e plataformas para publicação e divulgação de informações de saúde ao público em geral estão entre os principais exemplos de iniciativas na área de saúde digital que já vêm sendo implantadas em vários países. Ainda de acordo com a OMS (WORLD HEALTH ORGANIZATION, 2015), 78% dos países possuem alguma estratégia para regulação do Big Data no setor saúde.

Ao considerar a importância do uso do Big Data como ferramenta essencial para a implantação da saúde digital, foi possível categorizar em sete eixos as ações promovidas pelos países. Incluem-se:

1. Previsões de fluxos de pacientes

Em artigo da Forbes, Marr (2016) detalha como quatro hospitais que fazem parte da Assistance Publique-Hôpitaux, de Paris, têm usado dados de várias fontes para fazer previsões diárias e horárias de quantos pacientes estarão em cada hospital. Um dos principais conjuntos de dados é o equivalente a 10 anos de registros de admissões hospitalares, que os cientistas de dados analisaram usando técnicas de "análises de séries temporais". Essas análises permitiram aos pes-

quisadores ver padrões relevantes nas taxas de admissão. Então, eles poderiam usar *machine learning* para encontrar os algoritmos mais precisos que previam tendências futuras de admissão. Uma equipe extra pode ser redirecionada quando há expectativa de um número elevado de pacientes, reduzindo o tempo de espera e melhorando a qualidade de atendimento.

O artigo de Wong *et al* (2015) se concentra no uso de Big Data para analisar a utilização de ambulâncias em emergências. A partir dos históricos de localizações dos destinos, podem ser feitas previsões sobre o seu uso.

2. Registros eletrônicos de saúde (EHRs)

Os registros contemplam a digitalização dos dados dos pacientes, que inclui informações demográficas, histórico médico, alergias, resultados de exames laboratoriais etc. Os documentos são compartilhados através de sistemas de informação e podem estar disponíveis para provedores dos setores público e privado. Cada registro é composto de um arquivo modificável, o que significa que os médicos podem implementar mudanças ao longo do tempo sem burocracia e sem risco de duplicação de dados.

Os EHRs também podem acionar avisos e lembretes de quando um paciente deve fazer um novo teste de laboratório ou rastrear prescrições para ver se ele está seguindo as ordens dos médicos. A grande maioria dos hospitais americanos adotou os EHRs. O avanço possibilitou a criação de programas como o *PreManage ED,* no condado americano de Alameda, que compartilha registros de pacientes entre os departamentos de emergência. O compartilhamento permite que as equipes saibam se o paciente recebeu atendimento em outra unidade de saúde, quais exames foram realizados, quais os resultados e quais os tratamentos indicados. O programa auxilia para a não duplicação de exames (MEMON; KHOJA, 2020).

Em países em desenvolvimento, temos o exemplo da Índia. Desde 2010, o governo indiano tem emitido cartões *Aadhaar* e números de identificação exclusivos, associados à identificação biométrica, para todos os 1,2 bilhão de cidadãos. O cadastro possibilita a geração e o monitoramento de dados sociais e de saúde, incluindo registros médicos eletrônicos de famílias de baixa renda. Mesmo não totalmente implementado, o sistema permite uma coleta de dados mais confiável e auxilia em uma maior abrangência das estatísticas relacionadas à saúde.

3. Sistema de vigilância para controle e prevenção de doenças

Dispositivos inteligentes e computação em nuvem deram origem a novas abordagens de análise de Big Data de saúde pública. Um exemplo recente foi o uso de aplicativos de rastreamento da covid-19 usados em países da União Europeia, China, Coréia do Sul e Índia. Através da instalação de aplicativos nos telefones celulares foi possível a captação de dados dos indivíduos referentes à sua geolocalização e o monitoramento das condições de saúde. O acompanhamento auxiliou na construção de uma base de dados robusta, em tempo real, facilitando o controle sobre a disseminação do vírus.

Outro exemplo é a possibilidade de agrupamento de *coortes*, a partir da junção de várias bases de dados, cita sobre o caso dos Estados Unidos. A partir de dados disponibilizados dos seguros de saúde e receituários farmacêuticos foi possível identificar 742 fatores de riscos que predizem com alto grau de precisão se algum indivíduo está em risco de abusar do uso de opinóides (BRESNICK, 2018).

4. Ganhos nos diagnósticos por imagem

O uso combinado do Big Data em conjunto com *marching learning* (técnicas que se ajustam aos modelos algoritmicamente, adaptando-se aos padrões nos dados), promovem uma nova etapa da medicina de diagnósticos. A experiência de Sebastian Thrun, da Universidade de Stanford, proporcionou o aprendizado da inteligência artificial por meio de algoritmos a partir de uma rede neural de computação com 129 mil imagens de lesões da pele classificadas por dermatologistas (TROYANSKAYA *et al*., 2020). Em junho de 2015, Thrun e sua equipe começaram a validar o sistema usando um conjunto de 14 mil imagens que haviam sido diagnosticadas por dermatologistas, solicitando que o sistema reconhecesse três tipos de lesão: benignas, malignas e crescimentos não cancerosos (TROYANSKAYA *et al*., 2020). O sistema acertou 72% das vezes, comparado com um acerto de 66% obtido por dermatologistas qualificados. A experiência de Thrun foi ampliada para incluir 25 dermatologistas e uma amostra de 2 mil casos biopsiados. A máquina continuou sendo mais acurada do que o diagnóstico humano. Atualmente, o número de diagnósticos utilizando a técnica vem ganhando espaço nas práticas médicas (ILYASOVA *et al*., 2018; TROYANSKAYA *et al*., 2020).

5. Avanço nos tratamentos de doenças

O tratamento de doenças crônicas e degenerativas, como a diabetes, Aids e a esclerose lateral amiotrófica também ganha destaque. A concatenação e a compilação de uma ampla literatura médica podem

acelerar a descoberta de genes ligados a estas doenças bem como o avanço de tratamentos individualizados baseados no sequenciamento genético (KUMAR *et al*, 2015; LECLERC-MADLALA *et al*, 2017).

6. Alocação dos recursos

O uso de Big Data na área de saúde permite o planejamento estratégico graças a melhores informações sobre as motivações das pessoas. Os gerentes de atendimento podem analisar os resultados do *check-up* entre diferentes grupos demográficos e identificar quais fatores desencorajam as pessoas a aceitar tratamento.

A Universidade da Flórida fez uso do Google *Maps* e de dados de saúde pública para preparar mapas de calor com foco em várias questões, como crescimento populacional e doenças crônicas. Foi possível comparar esses dados com a disponibilidade de serviços médicos. Os *insights* recolhidos permitiram-lhes rever a sua estratégia de serviço e adicionar unidades de atendimento às áreas mais problemáticas.

7. Telemedicina

A prática está presente no mercado há mais de 40 anos, mas atualmente, com a chegada de videoconferências *online*, *smartphones*, dispositivos sem fio e tecnologias vestíveis, ela conseguiu entrar em plena expansão. O termo refere-se à prestação de serviços clínicos remotos usando as tecnologias. Incluem o uso das tecnologias de informação e comunicação para oferecer aos profissionais da atenção básica serviços como a realização de exames com emissão de laudos à distância (o chamado telediagnóstico), o esclarecimento, pela internet ou telefone, de dúvidas sobre procedimentos clínicos e questões relativas a processo de trabalho (a teleconsultoria) e ainda ações de formação à distância (tele-educação).

Os médicos usam a telemedicina para fornecer planos de tratamento personalizados e evitar hospitalização ou readmissão, além de ampliar o alcance a áreas mais remotas. Esse uso de análise de dados na área da saúde pode ser vinculado ao uso de análise preditiva. Ele permite que os médicos prevejam eventos médicos graves com antecedência e evitem uma piora das condições do paciente.

Portanto, a adoção da tecnologia pelos sistemas de saúde globais se mostra diversificada, destacando o cruzamento de diferentes bancos de dados, proporcionando uma *coorte* significativa, até a criação de sistemas de vigilância epidemiológica, avanço na precisão de diagnósticos, desenvolvimento da biotecnologia e medicina individualizada. O objetivo

comum é a busca por uma maior precisão e assertividade, além de projeções seguras que conduzam a uma redução dos custos dos sistemas.

Deve-se destacar que o país é um dos poucos a ter os dados de saúde pública catalogados e categorizados pelo Ministério da Saúde

Experiências brasileiras

O Brasil ainda é um campo próspero para o incremento de tecnologia. Deve-se destacar que o país é um dos poucos a ter os dados de saúde pública catalogados e categorizados pelo Ministério da Saúde. Contudo, ainda se apresenta incipiente no uso de tais ferramentas e faltam meios para a adoção de tecnologias inovadoras e modernas. No entanto, algumas experiências brasileiras com grande base de dados podem ser destacadas. A seguir, elas são elencadas por ordem cronológica.

2007 – Programa Nacional Telessaúde Brasil Redes

A Telessaúde, como componente da Estratégia *e*-Saúde (Saúde Digital) para o Brasil, tem como finalidade a expansão e melhoria da rede de serviços de saúde, sobretudo da Atenção Primária à Saúde (APS), e sua interação com os demais níveis de atenção fortalecendo as Redes de Atenção à Saúde (RAS) do SUS. O aparato engloba teleconsultoria, telediagnóstico, telemonitoramento, telerregulação e tele-educação.

Em 2015, o programa estava em funcionamento em 22 estados. Reunia cerca de 6.000 pontos de Telessaúde, instalados em Unidades Básicas de Saúde (UBSs) de 2.600 municípios, abrangendo 50 mil profissionais de equipes de atenção básica/saúde da família com possibilidade de acesso aos serviços (BRASIL, 2015).

Contudo, impasses dentro do próprio Conselho Federal de Medicina se fazem presentes. A Resolução 2.227/2018 (BRASIL, 2018), que regulamentava os atendimentos online por médicos brasileiros (telecirurgias e telediagnósticos), foi revogada, ficando subordinada a Resolução 1.643/2002.

2015 – Cartão Digital SUS

O *e*-Saúde engloba um conjunto de iniciativas que visam aperfeiçoar a plataforma digital como ferramenta de promoção e acesso do cidadão a serviços de saúde. A principal ação está relacionada com a criação de um Cartão Digital do SUS, um instrumento que visa facilitar o atendimento ao cidadão, agilizando o processo de marcação e agendamento de consultas e exames. Permite o cadastramento de alergias, telefones de emergência, calcula a massa corpórea e facilita o acompanhamento da pressão e da glicemia através de gráficos. Possibilita também a verificação de informações básicas do paciente (uso de medicamentos, dados sobre vacinas) num sistema próprio: o Sistema Cartão Nacional

de Saúde. Esse acesso geralmente é feito pelos próprios médicos ou profissionais da área de saúde que venham a atender o usuário, além de possibilitar traçar o diagnóstico e ofertar o tratamento mais adequado ao histórico do paciente.

2016 – Centro de Integração de Dados e Conhecimentos para a Saúde (Cidacs)

Inaugurado em dezembro de 2016 e localizado no Parque Tecnológico da Bahia, o Cidacs visa realizar estudos e pesquisas, desenvolver novas metodologias investigativas e promover capacitação profissional e científica, tendo por base projetos interdisciplinares fundados na integração de grandes bases de dados. Suas plataformas contam com: coorte de 100 milhões de brasileiros; Plataforma Zika; tecnologias e inovações para o SUS; equidade e sustentabilidade urbana; bioinformática e epidemiologia genética (Epigen). O uso de Big Data possibilita o direcionamento de políticas públicas no SUS.

Esta iniciativa pioneira associa a Fiocruz Bahia a uma série de importantes instituições científicas nacionais: Universidade Federal da Bahia (Instituto de Saúde Coletiva, Escola de Nutrição, Faculdade de Economia, Instituto de Física, Instituto de Matemática e Estatística), Senai-Cimatec, Fiocruz Brasília, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Icict Fiocruz/RJ), Universidade de Brasília e Fundação Getúlio Vargas. E internacionais: London School of Hygiene and Tropical Medicine e Farr Institute.

2017 – Registro Eletrônico em Saúde (RES)

O grande destaque do Programa de Informatização das Unidades Básicas de Saúde (PIUBS) fica a cargo da implementação do Registro Eletrônico em Saúde. É uma plataforma digital que permite o acompanhamento do histórico clínico do paciente em todas as Unidades Básicas de Saúde (UBSs), oferecendo ganho na qualidade e na gestão da atenção básica para o gestor, para os profissionais de saúde e para o cidadão.

Com a plataforma digital, toda a rede de saúde poderá acompanhar o histórico, os dados e os resultados de exames dos pacientes, verificar em tempo real a disponibilidade de medicamentos ou mesmo registrar as visitas de agentes de saúde, melhorando o atendimento ao cidadão. A transmissão 100% digital dos dados da rede municipal à base nacional permitirá, ainda, que o Ministério da Saúde acompanhe de forma mais próxima os gastos em saúde.

Segundo balanço sobre o aplicativo Meu DigiSUS, divulgado pelo Departamento de Monitoramento e Avaliação do SUS em 2018, de um total

de 42,5 mil UBSs distribuídas em mais de 5.500 municípios, 15,5 mil utilizam os prontuários eletrônicos (36%). A grande maioria das unidades, 27 mil, ainda não possui sistema de prontuário eletrônico. Das unidades que utilizam, cerca de 4 mil usavam versões oferecidas gratuitamente pelo Ministério da Saúde e 9 mil softwares próprios e privados. A dependência de programas privados se amplia à medida que se instala um processo de esvaziamento do Departamento de Informática do Ministério, o Datasus, e a criação do Departamento de Saúde Digital e avanços de toda infraestrutura (ANTUNES, 2019).

2018 – Meu DigiSUS

É uma plataforma móvel e digital disponibilizada pelo Ministério da Saúde para dar comodidade e autonomia aos usuários e dar agilidade aos serviços no SUS. Por meio dele, a população já pode acompanhar via celular suas consultas e exames ambulatoriais nas UBSs informatizadas; dispensação de medicamentos; visualização do histórico de suas solicitações; posição na fila do Sistema Nacional de Transplantes entre outras funcionalidades relacionadas à saúde pública.

Já foi realizado 1,2 milhão de *downloads* do Meu DigiSUS, entre *smartphones* com sistemas IOS e *Android*. Um dos principais benefícios do aplicativo é o melhor atendimento aos pacientes do SUS, onde eles poderão se tornar fiscais, avaliando o atendimento realizado, e denunciando fraudes em qualquer canto do País, além de possibilitar aos gestores municipais, estaduais e da União um planejamento adequado do setor, permitindo o aprimoramento constante desses serviços. A unificação dos serviços em uma única ferramenta também permitirá a correta aplicação dos recursos públicos.

Pela plataforma móvel oficial do SUS, o cidadão consegue encontrar hospitais, unidades de saúde e outros estabelecimentos próximos de sua residência; identificar farmácias participantes do Aqui tem Farmácia Popular e acompanhar os medicamentos que o cidadão retirou, além de avaliar o atendimento desses serviços. Também é possível acessar uma linha do tempo de cada atendimento realizado pelo SUS, além do Cartão Nacional de Saúde e os dados pessoais, com informações sobre nutrição e alergias.

Segundo o Ministério da Saúde, o aplicativo está em funcionamento há três anos e já é reconhecido pela sua inovação tecnológica. A plataforma é interligada às 19.788 Unidades Básicas em Saúde (UBS), que já estão informatizadas em 3.780 municípios, totalizando 106.179.196 pessoas cobertas (DATASUS, 2018). Ao todo, 11 sistemas estão integrados ao aplicativo, entre eles o Cadastro Nacional de Usuário do SUS (CADSUS), Cadastro Nacional de Estabelecimentos de Saúde (CNES), Farmácia Po-

É interessante notar que as iniciativas de uso das novas tecnologias voltadas para o SUS focalizam nas UBSs, que são as principais estruturas físicas da atenção básica

pular e os Sistemas Nacional de Transplantes (SNT), de Regulação (SISREG), de Atenção Básica (e-SUS AB) e o Hemovida (DATASUS, 2018).

É interessante notar que as iniciativas de uso das novas tecnologias voltadas para o SUS focalizam nas UBSs, que são as principais estruturas físicas da atenção básica. A atenção básica, ou primária, é conhecida como a porta de entrada dos usuários nos sistemas de saúde. Ou seja, é o atendimento inicial. Seu objetivo é orientar sobre a prevenção de doenças, solucionar os possíveis casos de agravo e direcionar os mais graves para níveis de atendimento superiores em complexidade. Assim, as UBSs desempenham um papel central na garantia de acesso a uma saúde de qualidade e possuem o potencial de diagnósticos precoces de doenças crônicas, o que facilita o controle das doenças em estágios iniciais e o não encaminhamento para as unidades de média e alta complexidades.

Observa-se que as iniciativas pontuadas ainda são descoordenadas, fragmentadas e pouco integradas. O avanço lento do uso do Big Data é representado pela baixa adesão das unidades básicas, pela falta de integração entre as redes de complexidades do SUS (básica, média e alta) e pelo incipiente compartilhamento entre o Sistema Único de Saúde e os planos complementares.

DESAFIOS PARA O AVANÇO DO USO DE BIG DATA NO SUS

Para que os profissionais de saúde, formuladores de políticas, pacientes e indivíduos tomem decisões mais assertivas sobre as questões sanitárias, se faz necessário ter o conhecimento do máximo de informações. O uso da tecnologia pode ser empregado na coleta e gerenciamento de dados. Nesse sentido, o uso de Big Data se torna viável, pois, diante a captação de informações individuais, é possível realizar análises e aplicar as informações de saúde obtidas.

Conforme as diretrizes da Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709, de 14 agosto de 2018) qualquer informação de dados pessoais é resguardada mantendo a privacidade da fonte de dados, uma vez que essa discussão perpassa por debates éticos, regulatórios e tecnológicos. Os principais são: i) a captação dos dados individuais, que devem ser robustos e a privacidade das informações deve ser garantida; ii) o estágio de governança deve ser reforçado, para garantir o respeito aos valores e princípios do uso dos dados; iii) o risco de violação acidental ou intencional da segurança do banco de dados deve ser minimizado pelas leis e por mecanismos tecnológicos adequados; iv) a presença de padrões de interoperabilidade que permitam que os dados sejam captados, agrupados e acessados sem maiores empecilhos; v) existência de

regras normativas que salvaguardam o uso e o compartilhamento dos dados em tempo real e vi) os dados disponibilizados devem estar formatados acessivelmente para o uso comum de pacientes, profissionais da saúde e formuladores de políticas (WYBER *et al*, 2015).

Para além dos desafios pontuados pelos autores, a ampliação do uso do Big Data pelo Sistema Único de Saúde é limitada pela falta de coordenação e recursos financeiros, dificuldades metodológicas, insuficiência e capacitação de recursos humanos. Boa parte pode ser atribuída à aprovação da Emenda Constitucional (EC), em dezembro de 2016. De acordo com o Conselho Nacional de Saúde (2020), com o chamado teto dos gastos públicos, o orçamento para a Saúde tem diminuído cada vez mais. Somente em 2019, a perda de investimentos na área representou R$ 20 bilhões, o que significa, na prática, a desvinculação do gasto mínimo de 15% da receita da União com a Saúde. Em 2017, quando a emenda passou a vigorar, os investimentos em serviços públicos de Saúde representavam 15,77% da arrecadação da União. Já em 2019, os recursos destinados à área representaram 13,54%. Frente ao desafio de subfinanciamento e pressões de demanda pelo serviço, a tendência é que haja um reforço na demora para a absorção de novas tecnologias pelo SUS.

Frente ao desafio de subfinanciamento e pressões de demanda pelo serviço, a tendência é que haja um reforço na demora para a absorção de novas tecnologias pelo SUS

CONCLUSÃO

À medida que as nações se desenvolvem ocorrem mudanças no padrão demográfico e epidemiológico; assim, aparecem outras formas de adoecer e morrer. Há a persistência de antigos problemas de saúde, aumento da relevância de doenças crônicas, queda de fecundidade e aumento da expectativa de vida. Todas essas modificações epidemiológicas influenciam na determinação dos novos objetivos do setor saúde, dado que agravam a complexidade da abrangência das políticas públicas. As políticas sociais se apresentam, nessa nova fase, muito mais complexas do que no passado e requerem conhecimentos aprofundados e capacidade de gestão que o setor público muitas vezes não dispõe.

O uso de novas abordagens tecnológicas sinaliza para ganhos significativos no que tange a redução de custos e melhorias de cobertura e qualidade na prestação de serviços públicos de saúde. A tecnologia em saúde tem a capacidade de mudar as configurações de alguns elementos técnicos e sociais, podendo coordenar aspectos clínicos e organizacionais do serviço de saúde. Isso ocorre através das relações dinâmicas e recíprocas entre interesses sociais, do sistema de saúde e dos indivíduos envolvidos. As tecnologias permeariam, dessa forma, o

O emprego de tecnologia [...] são respostas à crescente necessidade em aumentar a eficiência de todos os recursos empregados pelo setor público e à busca pela ampliação da cobertura de saúde conforme as alterações do perfil social vão acontecendo

ambiente dos pacientes, profissionais e da organização dos serviços em saúde e a adoção de novas tecnologias em saúde.

O emprego de tecnologia, a exemplo do Big Data, são respostas à crescente necessidade em aumentar a eficiência de todos os recursos empregados pelo setor público e à busca pela ampliação da cobertura de saúde conforme as alterações do perfil social vão acontecendo. Utilizar informações reais e atuais pode contribuir para reduzir a ineficiência na gestão público-privada, para buscar melhorias no gasto dos recursos e como sinalizador entre o que está sendo realizado e as metas definidas no plano gerencial. Isto é, aumentar a capacidade em fazer o uso mais adequado dos recursos conforme os objetivos pretendidos, para sanar as necessidades e os desejos dos agentes econômicos.

Ao melhorar a forma como as ações são planejadas e executadas, é possível, por meio do atendimento adequado às necessidades de saúde, aumentar o bem-estar geral de uma população, o que, por sua vez, contribui significativamente para o desempenho econômico do País. Assim, fortalecer os serviços de saúde com uso de tecnologias gera externalidades positivas, contribuindo com a maior eficiência do funcionamento do sistema econômico.

É inegável que o Brasil ainda precisa avançar no quesito tecnologia e uso de Big Data. Contudo, é um dos países que mais possuem dados de saúde relacionados, principalmente, à saúde pública. A experiência brasileira com uso de *Big Data* tem se mostrado positiva tanto para auxiliar o planejamento estratégico do setor saúde, como também na realização de políticas públicas, como verificado nos exemplos citados: o Programa Nacional Telessaúde Brasil Redes (2007), o Cartão Digital SUS (2015), o Centro de Integração de Dados e Conhecimentos para a Saúde (CIDACS – 2016), o Registro Eletrônico em Saúde (RES – 2017) e o Meu DigiSUS (2018). Essas e outras iniciativas visam à melhoria dos serviços de saúde, interligando a interface tecnológica, os serviços disponibilizados nos diferentes níveis de atenção à saúde, com as informações individuais dos pacientes, para que resulte no aumento da qualidade do serviço prestado e na melhoria da gestão dos recursos públicos.

Após esta breve discussão sobre a estrutura e desempenho do Sistema de Saúde do Brasil e o emprego de tecnologias inteligentes, é possível fazer alguns apontamentos. A interação entre os padrões sociais e a dinâmica econômica da saúde, juntamente com a incorporação de tecnologias estratégicas, contribui para que o País possa se inserir competitivamente no cenário mundial. Contudo, o Brasil carece de reorganizar a sua estratégia de articulação entre as áreas sociais, econômicas e tecnológicas para propiciar esse desenvolvimento e reduzir a sua vulnerabilidade e dependência externa. Cabe ao Estado, a participação efetiva

para subsidiar o desenvolvimento tecnológico e produtivo da saúde, incentivando, regulando e dando suporte à base produtiva do Sistema Nacional de Inovação em Saúde.

REFERÊNCIAS

ANTUNES, A. *Saúde digital: o que isso pode significar para o SUS?* Rio de Janeiro: EPSJV, Fiocruz, 2019. Disponível em: http://www.epsjv.fiocruz.br/noticias/reportagem/saude-digital-o-que-isso-pode-significar-para-o-sus. Acesso em: 15 maio 2020.

ARAÚJO, E. O Sistema de Saúde no Brasil. *In*: FÓRUM SAÚDE PÚBLICA X SAÚDE PRIVADA, 2014, Rio de Janeiro. *Trabalho apresentado* [...]. Rio de Janeiro: COPPE, ago. 2014. Disponível em: http://www.mobilizadores.org.br/wp-content/uploads/2014/08/O-sistema-de-saúde-brasileiro.pdf. Acesso em: 29 maio 2020.

ASSEMBLEIA MUNDIAL DA SAÚDE, 72., 2019, Genebra. *Anais* [...]. Genebra: OPAS, 17 maio 2019.

AZEVEDO, P. F. de. *et al*. . A cadeia de saúde suplementar no Brasil: avaliação de falhas de mercado e propostas de políticas. , São Paulo: INSPER Centro de Estudos em Negócios, maio, 2016. (White paper, n. 1). Disponível em: https://www.insper.edu.br/wp-content/uploads/2018/09/estudo-cadeia-de-saude-suplementar-Brasil.pdf. Acesso em: 18 maio 2020.

BACHA, E. L.; SCHWARTZMAN, S. *Brasil*: a nova agenda social. Rio de Janeiro: LCT, 2011.

BRASIL. HYPERLINK \hLei nº 8.080, de 19 de setembro de 1990. Dispõe sobre as condições para a promoção, proteção e recuperação da saúde, a organização e o funcionamento dos serviços correspondentes e dá outras providências. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 20 set. 1990.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 15 ago. 2018a.

BRASIL. Ministério da Saúde. *Estratégia e-Saúde para o Brasil*. Brasília: Ministério da Saúde, 2017. Disponível em: https://www.conasems.org.br/wp-content/uploads/2019/02/Estrategia-e-saude-para-o-Brasil-1.pdf. Acesso em: 1 jun. 2020.

BRASIL. Ministério da Saúde. Portal da Saúde. *DATASUS*: informações de saúde (TABNET). Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area= 0204&id=11671&VObj=http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/. Acesso em: 21 out. 2018b.

BRASIL. Ministério da Saúde. Portal da Saúde. *Sistema Único da Saúde*. Brasília: Ministério da Saúde. Disponível em: http://portalms.saude.gov.br/sistema-unico-de-saude/sistema-unico-de-saude. Acesso em: 15 maio 2020.

BRASIL. Ministério da Saúde. Portaria nº 1.101, de 12 de junho de 2002. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 13 jun. 2002. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2002/prt1101_12_06_2002.html. Acesso em: 17 maio 2020.

BRASIL. Ministério da Saúde. *Programa Nacional de Telessaúde Brasil Redes*. Brasília: Ministério da Saúde, 2015. Disponível em: http://bvsms.saude.gov.br/bvs/folder/programa_nacional_telessaude_bbrasil_redes_2015.pdf. Acesso em: 15 maio 2020.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. *Vigitel Brasil* 2011: prevalências de fatores de risco e de proteção: população adulta brasileira. Boletim Epidemiológico, Brasília, v. 44, n. 12, 2013.

BRESNICK, J.BCBSA Tackles opioids with Big Data, accreditations, member hotline. *Populations Health News*. Danvers, 12 July 2018. Disponível em: https://healthitanalytics.com/news/bcbsa-tackles-opioids-with-big-data-accreditations-member-hotline. Acesso em: 22 maio 2020.

CARVALHO, C. A. de; PINHO, J. R. O.; GARCIA, P. T. *Epidemiologia*: conceitos e aplicabilidade no Sistema Único de Saúde. São Luís: EDUFMA, 2017.

COMISIÓN ECONÓMICA PARA AMÉRICA LATINA Y EL CARIBE. *Big data and open data as sustainability tools*: a working paper prepared by the Economic Commission for Latin America and the Caribbean. Chile: CEPAL, out. 2014. Disponível em: https://repositorio.cepal.org/bitstream/handle/11362/37158/1/S1420677_en.pdf. Acesso em: 15 jul. 2019.

CONSELHO FEDERAL DE MEDICINA. Resolução nº 2.227, de 13 de dezembro de 2018. Define e disciplina a telemedicina como forma de prestação de serviços médicos mediados por tecnologias. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 6 fev. 2019.

CONSELHO NACIONAL DE SAÚDE (Brasil). *Saúde perdeu R$ 20 bilhões em 2019 por causa da EC 95/2016*. Brasília, 28 fev. 2020. Disponível em: https://conselho.saude.gov.br/ultimas-noticias-cns/1044-saude-perdeu-r-20-bilhoes--em-2019-por-causa-da-ec-95-2016#:~:text=Desde%20que%20a%20Emenda%20Constitucional,da%20Uni%C3%A3o%20com%20a%20Sa%C3%BAde. Acesso em: 15 maio 2020.

COSTA, L. S. Inovação nos serviços de saúde: apontamentos sobre os limites do conhecimento. *Cadernos de Saúde Pública*, Rio de Janeiro, v. 32, n 14, p. 1-12, 2016. Disponível em: https://www.arca.fiocruz.br/handle/icict/28099. Acesso em: 5 abr. 2019.

DATASUS. *Aplicativo do SUS aproxima cidadãos dos serviços públicos de saúde*. Disponível em: http://datasus1.saude.gov.br/noticias/atualizacoes/1142-aplicativo-do-sus-aproxima-cidadaos-dos-servicos-publicos-de-saude. Acesso em: 21 out. 2018.

FURTADO, C. *Dialética do desenvolvimento*. Rio de Janeiro: Fundo de Cultura, 1964.

GADELHA, C. Saúde e desenvolvimento: uma nova abordagem para uma nova política. *Revista Saúde Pública*, Rio de Janeiro, v. 46, p. 5-8, dez. 2012. Suplemento.

GADELHA, C. *et al*. . Saúde e desenvolvimento., *Informe CEIS*, Rio de Janeiro, v. 2, n. 2, dez. 2011.

HADOOP. *Big data analysis framework*. Estados Unidos: IBM , 2014.

ILYASOVA, N. *et al*. Particular use of big data in medical diagnostic tasks. *Pattern Recognition and Image Analysis*, Santiago de Compostela, v. 28, n. 1, p. 114–121, mar. 2018.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Projeção da população*. Disponível em: https://www.ibge.gov.br/apps/populacao/projecao/. Acesso em: 7 jun. 2020.

KUMAR, S. *et al*. Predictive methodology for diabetic data analysis in big data. *Procedia Computer Science*, Amsterdã, v. 50, p. 203-208, 2015.

LECLERC-MADLALA, S.; BROOMHALL, L.; FIENO, J. The 'end of AIDS' project: Mobilising evidence, bureaucracy, and big data for a final biomedical triumph over AIDS. *Global Public Health*, UK, v. 13, n. 8, p. 972-981, Dec. 2017.

MALIK, A. M. Oferta em serviços de saúde. *Revista US*, São Paulo, n. 51, p. 146-157, nov. 2001.

MARR, B. Big data in healthcare: Paris hospital predict admission rates using machine learning. *Forbes*, Estados Unidos, Dez. 2016. Disponível em: https://www.forbes.com/sites/bernardmarr/2016/12/13/big-data-in-healthcare-paris--hospitals-predict-admission-rates-using-machine-learning/#eff957c79a2c. Acesso em: 1 jun. 2020.

MATHIAS, M. *Antes do SUS*: como se (des)organizava a saúde no Brasil sob a ditadura. Rio de Janeiro, 10 abr. 2018. Disponível em: http://www.cee.fiocruz.br/?q=antes-do-sus. Acesso em: 13 ago. 2018.

MCKINSEY GLOBAL INSTITUTE. *Big data*: the next frontier for innovation, competition and productivity. Reino Unido: McKinsey & Company, May, 2011.

MEDICI, A. Propostas para melhorar a cobertura, a eficiência e a qualidade no setor de saúde. *In*: BACHA, E. L; SCHWARTZMAN, S. (org.). *Brasil*: A nova agenda social. Rio de Janeiro: LCT, 2011. p. 386.

MEMON, Q.; KHOJA, S. A. *Data Science*: theory, analysis, and applications. Estados Unidos: CRC Press, 2020.

MONITORAMENTO da assistência hospitalar no Brasil (2009-2017). *Boletim Informativo do PROADESS*, Rio de Janeiro, n. 4, fev. 2019. Disponível em: https://www.proadess.icict.fiocruz.br/Boletim_4_PROADESS_Monitoramento%20da%20assistencia%20hospitalar_errata_1403.pdf. Acesso em: 17 maio 2020.

PESQUISA NACIONAL DE SAÚDE 2013: acesso e utilização dos serviços de saúde, acidentes e violências: Brasil, grandes regiões e unidades da federação. Rio de Janeiro: IBGE, 2015. 100 p. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv94074.pdf. Acesso em: 17 maio 2020.

TROYANSKAYA, O.; TRAJANOSKI, Z.; CARPENTER, A.; THRUN, S.; RAZAVIAN, N.; OLIVER, N. Artificial intelligence and cancer. *Nature Cancer*, v. 1, p. 149-152, 2020.

VIACAVA, F. *et al*. SUS: oferta, acesso e utilização de serviços de saúde nos últimos 30 anos. *Ciência & Saúde Coletiva*, Rio de Janeiro, v. 23, n. 6, p. 1751-1762, jun. 2018.

WONG, H. T. *et al*. Big data as a new approach in emergency medicine research. *Journal of Acute Disease*, Haikou, v. 4, n. 3, p. 178–179, Aug. 2015.

WORLD HEALTH ORGANIZATION. *Third global survey on ehealth*. Suíça: WHO, 2015. Disponível em: https://www.who.int/goe/survey/2015survey/en/. Acesso em: 13 jun. 2020.

WORLD HEALTH ORGANIZATION. Universal health coverage. *Seventy-second World Health Assembly*. Suíça: WHO, Mar. 2019. Disponível em: https://www.who.int/about/governance/world-health-assembly/seventy-second-world-health-assembly. Acesso em: 21 maio 2020.

WYBER, R. *et al*. Big data in global health: improving health in low and middle-income countries. *Bulletin of the World Health Organisation*, Suíça, v. 93, n.3, p. 203-208, Mar. 2015.

ZUCCHI, P.; NERO, C. D.; MALIK, A. M. Gastos em saúde: os fatores que agem na demanda e na oferta dos serviços de saúde. *Saúde e Sociedade*, São Paulo, v. 9, n. 1-2, p. 127-150, 2000.

Resumo

Este trabalho visa caracterizar a homogeneidade dos Territórios de Identidade da Bahia a partir de dados multivariados agrupados em quatro aspectos: socioeconomia, condição do produtor rural, uso da terra e efetivo de animais. Foi utilizado o método de visualização de dados Planos de Componentes, baseado na Rede Neural Artificial não-supervisionada do tipo Mapa Auto-Organizável. A caracterização foi subsidiada pelo Índice de Homogeneidade Territorial calculado para todos os Territórios de Identidade e por estudos socioeconômicos da Superintendência de Estudos Econômicos e Sociais da Bahia. Os resultados confirmaram a heterogeneidade dos Territórios de Identidade Chapada Diamantina, Litoral Norte e Agreste Baiano e Portal do Sertão, assim como mostrou que territórios considerados homogêneos também apresentam dissimilaridades entre os seus municípios, como o Médio Sudoeste da Bahia.
Palavras-chave: Aprendizagem de máquina. Visualização de dados. Redes Neurais Artificiais. Bahia.

Abstract

This work aims to characterize the homogeneity of the Identity Territories of Bahia based on multivariate data grouped in four aspects: socioeconomics, rural producer status, land use and livestock. It has been applied the Component Planes data visualization method based on the unsupervised Self-Organizing Map Artificial Neural Network. The characterization has been supported by the Territorial Homogeneity Index calculated for all Identity Territories and by socioeconomic studies of the Superintendency of Economic and Social Studies of Bahia. The results confirmed the heterogeneity of the Chapada Diamantina, Litoral Norte and Agreste Baiano and Portal do Sertão Identity Territories, as well as shows that Territories considered homogeneous also present dissimilarities between their municipalities such as the Médio Sudoeste da Bahia.
Keywords: *Machine learning. Data visualization. Artificial Neural Network. Bahia.*

Caracterização da homogeneidade socioeconômica dos Territórios de Identidade a partir de Mapas Auto-Organizáveis

MARCOS AURÉLIO SANTOS DA SILVA
Doutor em Computação, pela Université Toulouse 1 Capitole e mestre em Computação Aplicada, pelo Instituto Nacional de Pesquisas Espaciais (INPE). Pesquisador da Embrapa Tabuleiros Costeiros.
marcos.santos-silva@embrapa.br

AS POLÍTICAS PÚBLICAS de promoção do desenvolvimento regional a partir do conceito de território ganharam força nas últimas décadas no Brasil. De fato, o processo de territorialização se associou à regionalização para criar instâncias de unidades espaciais que otimizassem a intervenção pública, mas que, sobretudo, possibilitassem a agregação de valor à região a partir da promoção de diferenciais locais (SABOURIN, 2015; SAQUET, 2010).

Enquanto a regionalização privilegia a identificação de regiões homogêneas a partir de questões e métricas puramente objetivas, como a proximidade física, características da paisagem geográfica ou atividades econômicas, a territorialização se atém a questões de identidade, cultura e pertencimento (VELLOSO, 2013). No Brasil destacam-se políticas territoriais nos âmbitos federal, como os Territórios Rurais e Territórios da Cidadania, e estadual, como os Territórios de Planejamento de Sergipe e Pará e Territórios de Identidade da Bahia.

30 municípios [registraram o desejo] de migrar de TI, o que denota a heterogeneidade intraterritorial e a necessidade de revisão constante dos processos que determinaram a composição dos TIs

Os Territórios de Identidade (TI) foram criados em 2010 pelo Decreto Estadual nº 12.354, que define TI em seu Art. 1º §1º como:

> [...] o agrupamento identitário municipal formado de acordo com critérios sociais, culturais, econômicos e geográficos e reconhecido pela sua população como o espaço historicamente construído ao qual pertence, com identidade que amplia as possibilidades de coesão social e territorial (BAHIA, 2010).

Os 27 TIs instituídos são formados por municípios contíguos e considerados homogêneos segundo critérios multidimensionais (SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAIS DA BAHIA, 2015, 2016, 2018; BLATT; GONDIM, 2013). Como afirma Superintendência de Estudos Econômicos e Sociais da Bahia (2015, p. 122) para o TI Chapada Diamantina: "No território observa-se um comportamento homogêneo entre os municípios em referência ao desempenho econômico e à estrutura social". No entanto, como observaram Santos, Silva e Pereira (2011), a partir de um estudo sobre a tipologia dos municípios baianos segundo a distribuição do valor agregado ao PIB por categoria, nenhum TI é perfeitamente homogêneo. De fato, Silva e Souza (2018) calcularam a homogeneidade dos TIs da Bahia a partir de 45 variáveis e constataram que há diferenças entre eles, sendo o TI Chapada Diamantina considerado o mais heterogêneo dentre todos.

Borges e Serpa (2012) registraram o desejo de 30 municípios de migrar de TI, o que denota a heterogeneidade intraterritorial e a necessidade de revisão constante dos processos que determinaram a composição dos TIs baseados em fatores socioculturais, de identidade e de pertencimento. Por exemplo, Figueiredo e outros (2018) destacam as desigualdades internas no TI Sudoeste Baiano, onde o crescimento econômico concentra-se quase que unicamente no município Vitória da Conquista.

Portanto, a identificação dos fatores que contribuem para a homogeneidade ou heterogeneidade dos Territórios de Identidade é muito importante para o gerenciamento desta política pública. Ao definir o TI a partir de critérios multivariados, incluindo principalmente fatores socioculturais, como destacam Velloso (2013) e Blatt e Gondim (2013), a política pública considera os aspectos e processos sociais complexos de cada região. Assim, capturar essa complexidade se apresenta como um desafio para os gestores públicos responsáveis por esta política.

Uma forma de descrever essa complexidade é através da análise de dados multivariados e oriundos de diversas fontes, incluindo dados massivos. Os dados massivos (Big Data) são definidos por seu volume, sua disponibilidade e pela dificuldade de tratamento, seja pelo volume dos dados a serem processados ou pela dificuldade de análise de dados

multivariados de diversos formatos e fontes. Neste último caso podemos incluir os dados administrativos agregados por município, oriundos de censos, pesquisas anuais etc. No caso dos TIs, é relevante calcular sua homogeneidade a partir deste tipo de dado agregado por município.

O objetivo deste estudo é caracterizar a homogeneidade dos TIs da Bahia a partir dos dados usados por Silva e Souza (2018) utilizando o método de visualização de dados Plano de Componentes, baseado na Rede Neural Artificial (RNA) do tipo Mapa Auto-Organizável (SOM – *Self-Organizing Map*) (KOHONEN, 2013, 2001). Esta técnica vem sendo aplicada em problemas semelhantes como a tipologia dos Territórios de Planejamento de Sergipe (SILVA *et al.*, 2011) e cálculo do Índice de Homogeneidade Territorial (IHT), aplicado aos Territórios de Identidade (SILVA; SOUZA, 2018). Na seção 2, são descritos os dados e o método de análise; na seção 3, são apresentados os resultados e discussões e na seção 4, as conclusões.

MATERIAL E MÉTODOS

Para caracterização da homogeneidade dos TIs da Bahia será analisado um conjunto de variáveis agrupadas em quatro aspectos utilizados por Silva e Souza (2018), a saber: socioeconômico, condição do produtor rural, uso da terra e efetivo de animais. De fato, procurou-se incluir na análise variáveis que representassem os elementos do ambiente rural que respondem por parte relevante das atividades econômicas dos municípios da Bahia.

A partir desses dados, Silva e Souza (2018) classificaram os TIs segundo o Índice de Homogeneidade Territorial (IHT) proposto pelos autores. Esta classificação será usada como ponto de referência na caracterização da homogeneidade dos TIs. Para a caracterização será utilizada a Rede Neural Artificial do tipo Mapa Auto-Organizável de Kohonen (2001, 2013), com aprendizagem de máquina não supervisionada em conjunto com a técnica de visualização de dados Planos de Componentes.

Área de estudo e dados

Para caracterizar a homogeneidade dos TIs foram utilizadas 45 variáveis organizadas em quatro aspectos, sendo 22 variáveis para o aspecto socioeconômico, incluindo dados do Programa Bolsa Família (INSTITUTO DE PESQUISA ECONÔMICA APLICADA, 2007) e do atlas do desenvolvimento humano (PROGRAMA DAS NAÇÕES UNIDAS PARA DESENVOLVIMENTO, 2010), cinco para o aspecto condição do produtor rural, 15 para o aspecto uso da terra e três para o aspecto efetivo de animais, todas estas oriundas do Censo Agropecuário (2006).

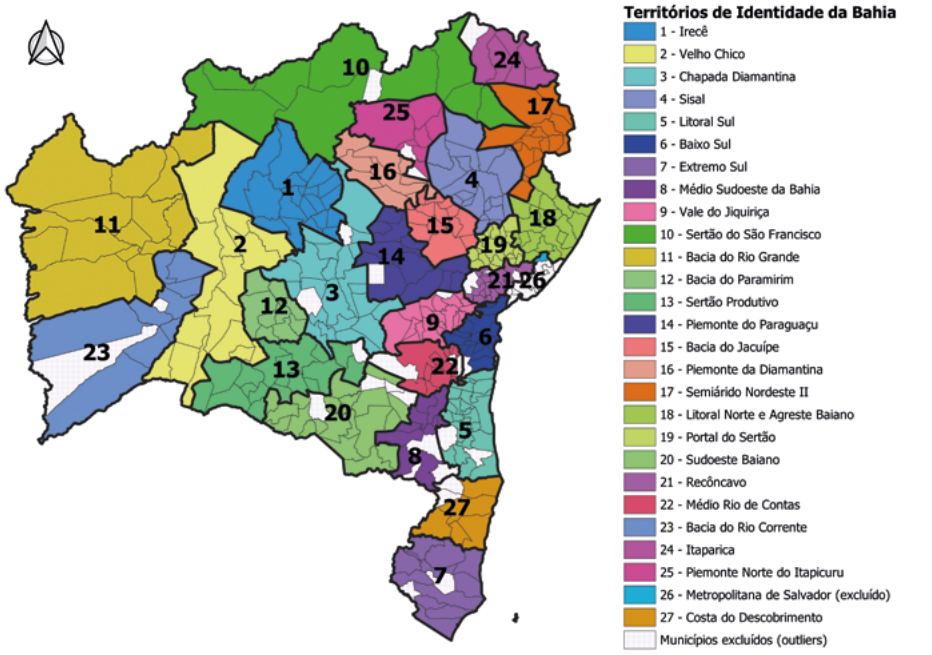
As variáveis abrangem indicadores de renda, pobreza, desigualdade social, educação, qualidade dos domicílios, distribuição demográfica entre os espaços rurais e urbanos, condição do produtor rural (proprietário, arrendatário, parceiro, ocupante ou produtor sem área), uso da terra (e.g., lavouras permanentes/temporárias, pastagens, matas e/ou florestas), efetivos de animais (bovinos, caprinos e ovinos) e produção de leite de vaca (ver Tabela A1 no Anexo A). Todas as variáveis foram padronizadas em função da média e do desvio padrão.

Importante destacar que Silva e Souza (2018) excluíram da análise alguns municípios que apresentaram valores acima de cinco desvios-padrão para alguma variável. Em função do número de municípios eliminados pelo critério anterior, o TI Metropolitana de Salvador também foi retirada da análise. Os dados usados neste estudo estão disponíveis em Silva (2019).

O mapa da Bahia e seus TIs da Figura 1 mostra os municípios eliminados do cálculo do IHT e da caracterização da homogeneidade dos TIs realizada neste trabalho.

Na Figura 2 tem-se o IHT médio calculado por Silva e Souza (2018) para os 26 TIs considerados, em ordem crescente do indicador. Ou seja, do mais heterogêneo (Chapada Diamantina) ao mais homogêneo (Médio

Figura 1
A área de estudo compreende 26 Territórios de Identidade da Bahia e 375 municípios, excluindo aqueles que apresentaram valores atípicos para as variáveis selecionadas. Também foi excluído o TI Metropolitana de Salvador

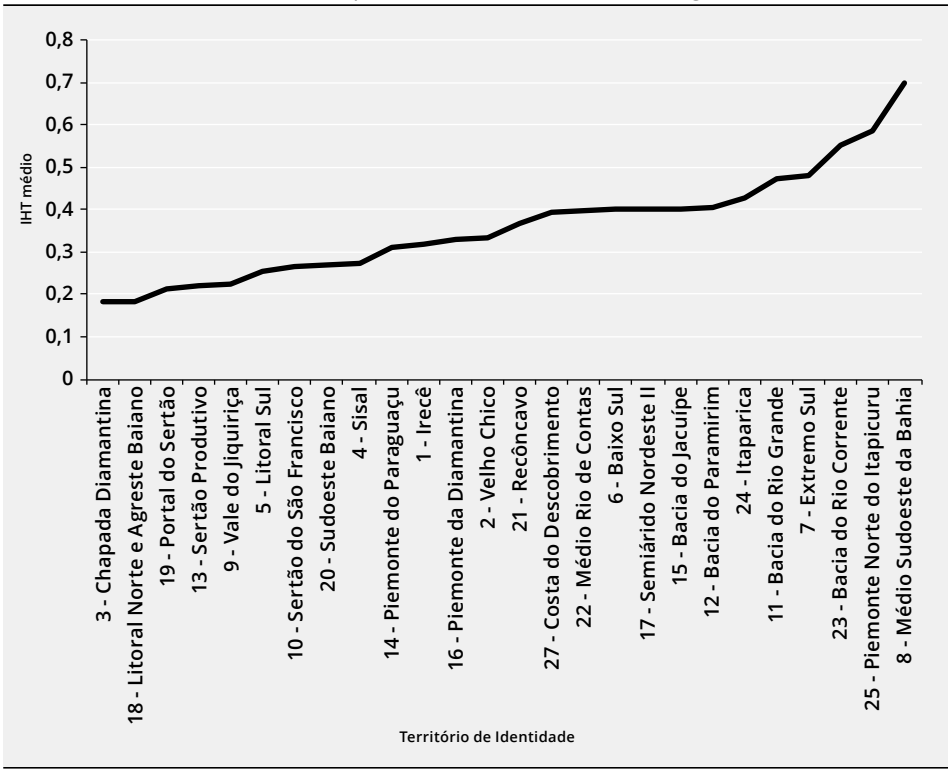


Fonte: Silva e Souza (2018).

Sudoeste da Bahia). Importante destacar que o IHT leva em consideração a área de cada município no TI. Assim, municípios com grande área em relação ao TI terão maior impacto no IHT, o que não será o caso da caracterização a partir da RNA SOM e dos Planos de Componentes.

Mais importante que o valor IHT para cada TI, é a posição relativa de cada um no gráfico

Figura 2
Gráfico da curva da média dos IHT para cada Território de Identidade do estado da Bahia, considerando todos os experimentos realizados com o algoritmo TerraSOM



Fonte: Silva e Souza (2018).

Mais importante que o valor IHT para cada TI, é a posição relativa de cada um no gráfico. Entre os valores 0,0 e 0,4 temos TIs com menor homogeneidade que aqueles entre 0,4 e 0,7, sendo que há um conjunto de seis municípios com IHT muito próximos de 0,4. A aceleração da curva a partir do TI Bacia do Paramirim sugere que os sete TIs mais homogêneos se destacam dos demais.

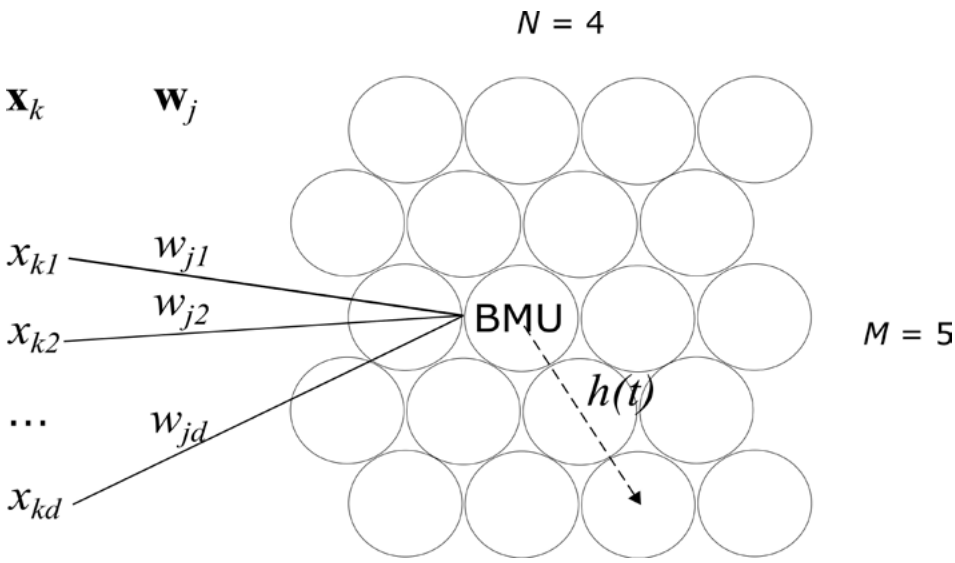
Visualização de dados com Mapas Auto-Organizáveis

A Rede Neural Artificial (RNA) do tipo Mapa Auto-Organizável de Kohonen (SOM – *Self-Organizing Map*) é um algoritmo de aprendizagem de máquina não-supervisionado que projeta os vetores de entrada x_k (conjunto de dados dos municípios) numa grade (em geral bidimensional com M linhas e N colunas) de neurônios artificiais, representados por vetores de referências w_j. A aprendizagem de máquina consiste num

Neste trabalho optou-se pela aprendizagem padrão, onde os vetores de entrada são apresentados aleatoriamente à RNA SOM e os vetores de referência são atualizados assim que é definido o BMU.

processo iterativo (número pré-definido de ciclos de apresentação dos dados de entrada à RNA SOM) onde cada vetor de entrada é apresentado à RNA SOM, que busca o vetor de referência mais próximo em função de uma medida de similaridade (neste trabalho, a distância euclidiana), após achar o neurônio artificial mais próximo (BMU – *Best Match Unit*) do vetor de entrada, o vetor de referência do BMU e de seus vizinhos na grade neural são atualizados de forma a se aproximarem do vetor de entrada. Uma função de vizinhança *h(t)* gaussiana centrada no BMU define quais neurônios artificiais serão atualizados, e uma função de aprendizagem *α(t)* determina a intensidade dessa aproximação (KOHONEN, 2001, 2013; SILVA *et al.*, 2011; SILVA *et al.*, 2015).

Figura 3
Exemplo de um Mapa Auto-Organizável bidimensional NxM, com entrada x_k e vetores de referência w_j



Fonte: Elaborado pelo autor.

Há diferentes estratégias de aprendizagem de máquina para a RNA SOM. Neste trabalho optou-se pela aprendizagem padrão, onde os vetores de entrada são apresentados aleatoriamente à RNA SOM e os vetores de referência são atualizados assim que é definido o BMU. Assim, seja Ξ o conjunto dos vetores de entrada composto por x_k, $k = 1, ..., n$, sendo n o número total de observações, tem-se o algoritmo de aprendizagem padrão ou sequencial, como segue:

a. Os vetores de referência, $w_j = [w_{j1}, w_{j2},...,w_{jp}]^T$, são iniciados linearmente a partir dos dois autovetores com maiores autovalores calculados a partir de Ξ.

b. Para cada tempo discreto *t*

1. Para todo $x_k \in \Xi$, *k = 1, ..., n*, encontre o neurônio vencedor *c* segundo a distância euclidiana:

$$c = \text{argmin}_j \left\{ \left\| x_k - w_j \right\| \right\}, j = 1,2,...,m \quad (1)$$

onde *m* corresponde ao número de neurônios na rede. A ordem de apresentação dos vetores de entrada deve ser aleatória.

2. Os vetores de código w_j do neurônio vencedor e dos seus vizinhos são, então, atualizados segundo a equação:

$$w_{ji}(t+1) = w_{ji}(t) + \alpha(t)h(t)[x_{ik}(t) - w_{ij}(t)] \quad (2)$$

onde $\alpha(t)$ é uma função que determina a taxa de aprendizagem na iteração *t* e *h(t)* é a função que determina a vizinhança entre o neurônio vencedor *c* e seus vizinhos.

Ao final do processo de aprendizagem cada vetor de entrada estará associado a um neurônio artificial, sendo que vetores de entrada similares estarão próximos na grade neural

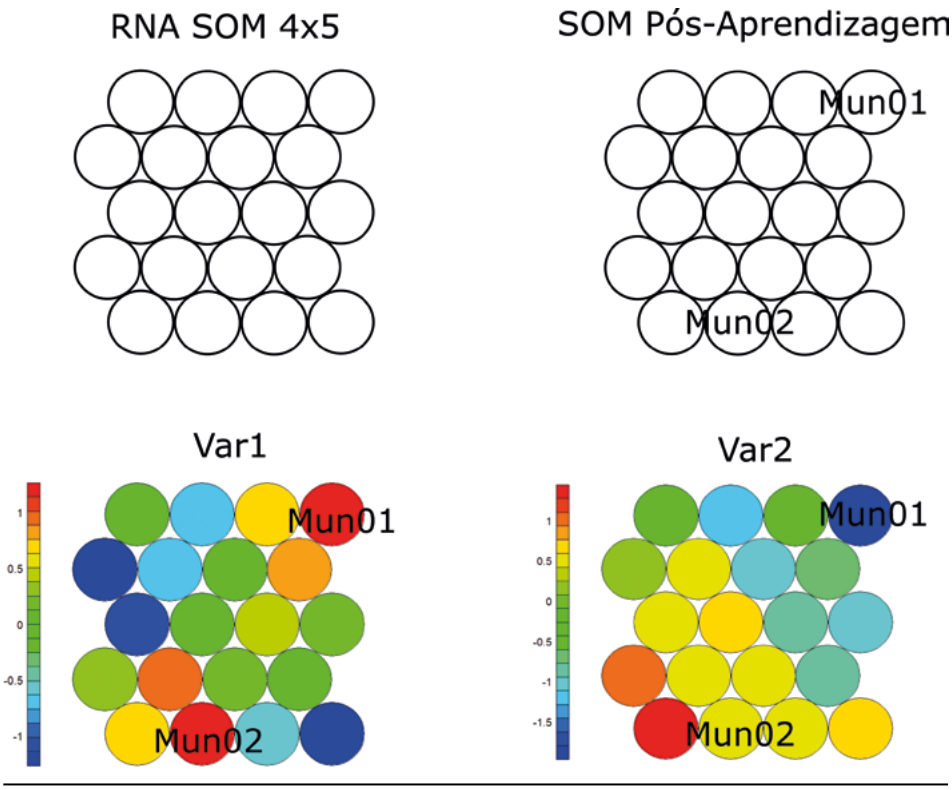
Ao final do processo de aprendizagem cada vetor de entrada estará associado a um neurônio artificial, sendo que vetores de entrada similares estarão próximos na grade neural. De fato, o processo de aprendizagem de máquina ordena topologicamente os dados de entrada na grade neural. Como cada neurônio artificial pode estar associado a mais de um vetor de entrada, a RNA SOM também pode ser usada (caso tenhamos mais dados de entrada que de vetores de referência) como compressor dos dados, mantendo as propriedades estatísticas dos dados de entrada. No entanto, é importante destacar que todos os vetores de entrada similares estarão próximos na grade neural, mas que nem todos os vetores de entrada associados a neurônios artificiais próximos serão similares entre si. Por isso, é necessário o uso de métodos de visualização de dados e análise de agrupamentos, em conjunto com o algoritmo de aprendizagem de máquina da SOM, para poder analisar os dados.

Um método bastante usado para análise da contribuição de cada variável do vetor de entrada na formação da configuração final da distribuição das observações na grade neural são os Planos de Componentes construídos a partir dos vetores de referência dos neurônios artificiais da grade neural. Assim, para cada variável será gerado um Plano de Componente, que é uma representação gráfica da distribuição dos valores de referências para aquela variável na grade neural. Ou seja, para cada neurônio observa-se o valor do vetor de referência relativo à variável em questão e se atribui uma cor para este neurônio. Como os dados de entrada estão padronizados usaremos um padrão de cor divergente, onde valores negativos (abaixo da média) estarão associados a tonalidades crescentes da cor vermelha e valor positivos (acima da média) a

tonalidades crescentes da cor azul. Na Figura 4 tem-se um exemplo de uma RNA SOM 4x5 que após o processo de aprendizagem de máquina associou cada observação a um neurônio. Neste caso tem-se duas variáveis e, respectivamente, dois Planos de Componente (Var1 e Var2).

Interpreta-se os Planos de Componente da seguinte forma: localiza-se a observação na grade neural, e.g. a observação "Mun01" no canto superior direito da grade neural; e analisa-se esta observação nos Planos de Componente em função do padrão de cor divergente. No exemplo tem-se que a observação "Mun01" apresenta valor muito acima da média para a variável "Var1" e muito abaixo para a variável "Var2".

Figura 4
Processo de construção dos Planos de Componente a partir de RNA SOM 4x5. Após o processo de treinamento a RNA é rotulada, associando um neurônio para cada vetor de entrada (e.g., Mun01 e Mun02). Para cada variável (e.g., Var1 e Var2) é associado um padrão de cor, no caso divergente, de acordo com o valor da variável no vetor de referência da RNA



Fonte: Elaborado pelo autor.

A caracterização da homogeneidade dos TIs seguirá as seguintes etapas: 1) define-se uma RNA de tamanho NxM e aplica-se o algoritmo de aprendizagem padrão; 2) para cada TI serão projetados na grade neural os seus municípios componentes observando se os mesmos estão agrupados ou não numa mesma região da grade neural. Quanto mais próximos os municípios estiverem na grade neural, mais homogêneo o

TI será; 3) serão selecionados os TIs com maior agregação dos seus municípios na grade neural de forma que seja possível a caracterização da sua homogeneidade; 4) serão gerados os Planos de Componentes para todas as 45 variáveis do vetor de referência da grade neural; 5) serão selecionados os Planos de Componentes que melhor caracterizam os TIs selecionados na etapa (3); 6) os TIs considerados homogêneos serão caracterizados em função dos Planos de Componentes selecionados.

Os estudos da Superintendência de Estudos Econômicos e Sociais da Bahia (2015, 2016, 2018) e os valores médios dos Índices de Homogeneidade Territorial (SILVA; SOUZA, 2018) foram utilizados como suporte em todas as etapas de caracterização dos TIs. Foi utilizado o pacote R *Kohonen* (WEHRENS; KRUISSELBRINK, 2018; WEHRENS; BUYDENS, 2007) para geração da grade neural após o processo de aprendizagem de máquina e geração dos Planos de Componentes.

RESULTADOS E DISCUSSÃO

Os resultados foram gerados por uma RNA SOM com 22 linhas e 17 colunas, totalizando 374 neurônios. Neste caso optou-se por um número total de neurônios (22x17=374) próximo do número total de observações (375). Esta RNA foi definida com a vizinhança entre os neurônios com formato hexagonal (6-vizinhos), o que aumenta o número de vizinhos para cada neurônio e consequentemente melhora o processo de quantização vetorial. Definiu-se a função alpha, valor com máximo igual a 0,05 e mínimo igual a 0,01, de acordo com Kohonen (2001) e a função de vizinhança gaussiana, que determina a intensidade com que cada vetor de referência vizinho será aproximado ao vetor de referência do BMU. Neste trabalho utilizou-se a aprendizagem de máquina padrão descrita na seção 2.2 com 500 ciclos de aprendizagem de máquina.

Conforme descrito na seção material e métodos foram elaboradas as projeções dos municípios na grade neural para todos os 26 TIs estudados (ver anexo B). No entanto, foram selecionados os seis onde houve agregação dos municípios componentes do TI numa determinada região da grade neural e três TIs onde houve dispersão dos municípios componentes. Neste último caso foram selecionados os três TIs que também apresentam os menores IHT conforme Silva e Souza (2018).

A caracterização da homogeneidade foi realizada para os seis TIs com maior agregação dos seus municípios componentes na grade neural, são eles: Baixo Sul, Extremo Sul, Médio Sudoeste da Bahia, Bacia do Jacuípe, Bacia do Rio Corrente e Piemonte Norte do Itapicuru. Para facilitar a interpretação dos Planos de Componentes para cada TI, foi definido o contorno do maior agrupamento de municípios do TI na grade neural.

Todas as variáveis contribuem para a formação do agregado dos municípios na grade neural, mas algumas mostram as diferenças entre os TIs mais explicitamente que outras

Foram gerados os Planos de Componentes para todas as variáveis (ver anexo C). No entanto, foram selecionadas nove variáveis que melhor evidenciam as diferenças entre os seis TIs selecionados. Destaca-se que todas as variáveis contribuem para a formação do agregado dos municípios na grade neural, mas algumas mostram as diferenças entre os TIs mais explicitamente que outras.

Na seção 3.1 discute-se a distribuição dos municípios na grade neural por TI estabelecendo um comparativo com o Coeficiente de Variação de indicadores socioeconômicos utilizados pela SEI para análise da homogeneidade dos TIs, na seção 3.2, a segmentação da RNA SOM separando os TIs na grade neural, e na seção 3.3 é realizada a caracterização dos seis TIs considerados homogêneos pela agregação dos seus municípios na grade neural.

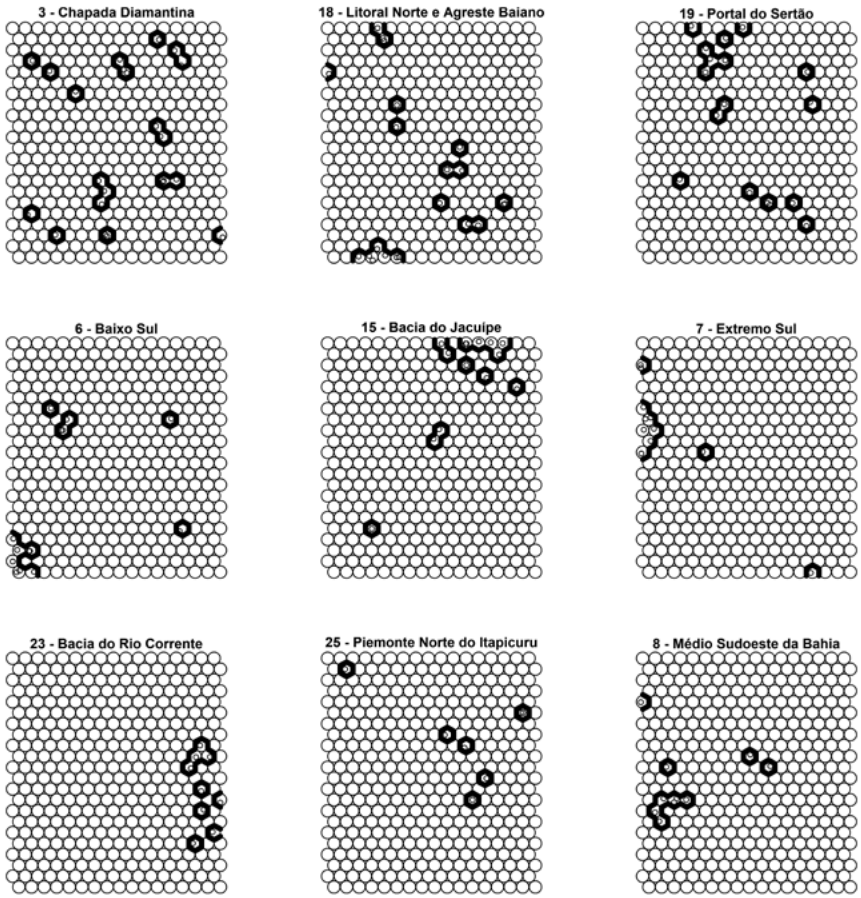
Distribuição dos municípios na grade neural

Na Figura 5 tem-se a projeção dos municípios na grade neural para nove dos 26 TIs analisados (todas as projeções encontram-se no anexo B). Estas projeções correspondem a dois grupos, as três primeiras correspondem aos municípios com menor IHT (Chapada Diamantina, Litoral Norte e Agreste Baiano e Portal do Sertão) e os demais a municípios com elevado IHT e que apresentaram algum grau de agregação dos municípios na grade neural (Baixo Sul, Bacia do Jacuípe, Extremo Sul, Bacia do Rio Corrente, Piemonte Norte do Itapicuru e Médio Sudoeste da Bahia).

Observa-se claramente que os três TIs com menor IHT apresentam seus municípios componentes dispersos na grade neural. Isto denota que, considerando as 45 variáveis selecionadas, há alta heterogeneidade destes TIs. Mesmo tendo sido excluídos os dois municípios (Bonito e Piatã) com valores atípicos para as variáveis escolhidas, o TI Chapada Diamantina apresenta alto grau de dispersão dos seus municípios na grade neural. A dispersão dos municípios Litoral Norte e Agreste Baiano confirma o efeito dos diferentes perfis geofísicos e climáticos destacado em em Superintendência de Estudos Econômicos e Sociais da Bahia (2016). Para o TI Portal do Sertão, tem-se forte dispersão, apesar da exclusão do município Feira de Santana, que concentra as principais atividades econômicas e 68% da população, conforme o Censo Demográfico (2012).

Na Tabela 1 tem-se o Coeficiente de Variação (CV) para cinco indicadores socioeconômicos utilizados pela SEI para caracterizar os nove TIs aqui tratados (SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAS DA BAHIA, 2015, 2016, 2018). Observa-se, em negrito, que os maiores CVs são observados justamente para os TIs com dispersão na grade neural e com baixo IHTs médios. Destaca-se o CV de 2.917% da taxa de crescimento populacional entre 2000 e 2010 do TI Chapada Diamantina.

Figura 5
Projeção de seis TIs representantes daqueles com maior homogeneidade (Baixo Sul, Bacia do Jacuípe, Extremo Sul, Bacia do Rio Corrente, Piemonte Norte do Itapicuru e Médio Sudoeste da Bahia) e três TIs com os menores valores de homogeneidade territorial segundo o índice IHT (Chapada Diamantina, Litoral Norte e Agreste Baiano e Portal do Sertão)



Fonte: Elaborado pelo autor.

Apesar de haver diferenças nos valores dos CVs para o indicador variação do IDH entre 1991 e 2010 observa-se a menor diferença entre o TI com maior CV (TI Portal do Sertão) e o menor (TI Médio Sudoeste da Bahia).

Os outros seis TIs mostram algum grau de especialização da grade neural num conjunto específico de municípios, sendo o TI Bacia do Rio Corrente aquele onde todos os nove municípios (foram excluídos da análise Jaborandi e Canápolis) estão agrupados numa mesma área da grade neural. No entanto, há aquelas com alguns municípios dispersos como o Baixo Sul com apenas 57% dos seus 14 municípios (foi excluído Ibirapitanga) agregados na parte inferior esquerda da grade neural. De fato, muito dificilmente todos os municípios do TI estarão agrupados numa região específica da grade neural, mesmo considerando que os municípios muito

Tabela 1
Coeficientes de variação para oito variáveis socioeconômicas utilizadas no estudo de elaboração dos perfis dos TIs

Coeficientes de Variação (CV)	3 - Chapada Diamantina	18 - Litoral Norte e Agreste Baiano	19 - Portal do Sertão	6 - Baixo Sul	15 - Bacia do Jacuípe	7 - Extremo Sul	23 - Bacia do Rio Corrente	25 - Piemonte Norte do Itapicuru	8 - Médio Sudoeste da Bahia
CV do Valor Adicionado Bruto ao PIB da Agropecuária em 2012 (%).	187,8	176,0	71,3	54,2	92,3	64,5	161,8	73,8	48,1
CV de receita própria em 2012 (%)	49,5	89,8	102,7	83,6	35,9	55,0	40,2	45,5	58,6
CV da taxa de crescimento populacional entre 2000 e 2010 (%)	2917,4	66,7	212,2	152,3	1088,4	220,3	1124,8	301,3	-1024,6
CV do percentual de pessoas sem ocupação em relação à população total do município em 2010 (%)	58,1	148,4	258,6	126,9	85,9	126,8	44,2	91,3	84,6
CV da variação do IDH entre os anos de 1991 e 2010 (%)	29,4	22,6	31,7	28,7	19,2	30,3	22,0	25,2	18,0

Fonte: Superintendência de Estudos Econômicos e Sociais da Bahia (2015, 2016, 2018).

distintos dos demais já foram excluídos da análise. Importante destacar que o TI Bacia do Rio Corrente é o mais agregado, mas não com maior IHT.

O TI com maior IHT, o Médio Sudoeste da Bahia, não apresenta os 10 municípios (foram excluídos Potiraguá, Maiquinique e Itapetinga) perfeitamente agregados na grade neural e isto pode representar variações internas no TI. De fato, apesar de apresentar os menores CVs para os indicadores VAB agropecuário e variação do IDH, mostrou forte CV para o crescimento populacional. De acordo com a Superintendência de Estudos Econômicos e Sociais da Bahia (2015) este TI apresenta algumas características comuns a todos os seus municípios componentes, que o torna homogêneo, como o número reduzido de habitantes, alto nível de urbanização e homogeneidade da participação dos diferentes setores no VAB.

Os TIs Piemonte Norte do Itapicuru (foram excluídos Pindobaçu e Caldeirão Grande) e Extremo Sul (foram excluídos Teixeira de Freitas e Lajedão) apresentaram altos valores para o IHT médio e apresentaram pouca dispersão dos seus municípios na grade neural. De fato, verifica-se que os CV de todos os indicadores da Tabela 1 dos dois TIs encontram-se em patamares intermediários. Segundo a Superintendência de Estudos Econômicos e Sociais da Bahia (2015), o TI Extremo Sul se destaca pelo alto nível de urbanização, forte presença da indústria de papel e celulose e produção agrícola de cana-de-açúcar e café. O TI

Piemonte Norte do Itapicuru se destaca pelo baixo nível de urbanização e presença da indústria extrativista mineral (SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAIS DA BAHIA, 2018).

Apesar da boa agregação dos municípios do TI Bacia do Jacuípe (foi excluído Capim Grosso) na grade neural, possui o oitavo IHT médio mais elevado. Segundo a Superintendência de Estudos Econômicos e Sociais da Bahia (2016) este TI se destaca pelo alto nível de analfabetismo e pobreza, baixo nível de urbanização e predomínio do setor terciário nas economias municipais.

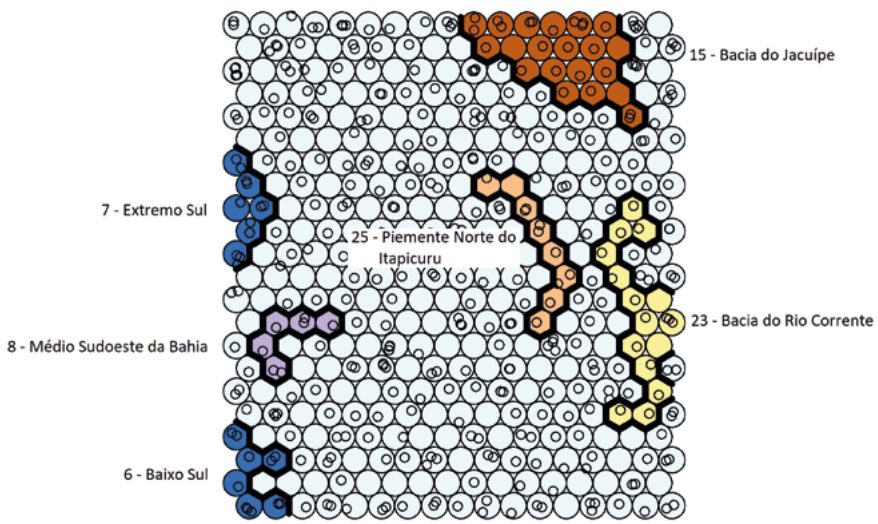
Segmentação da RNA SOM

A dispersão dos municípios dos TIs Chapada Diamantina, Litoral Norte e Agreste Baiano e Portal do Sertão dificulta a interpretação dos Planos de Componentes, cuja função é identificar justamente as correspondências entre diferentes regiões da grade neural e os Planos de Componentes. Assim, analisar os Planos de Componentes de TIs com municípios dispersos na grade neural pode ser uma tarefa difícil. Para resolver este problema decidiu-se considerar homogêneos os TIs com pelo menos 50% dos seus municípios agrupados numa região específica da grade neural.

Analisar os Planos de Componentes de TIs com municípios dispersos na grade neural pode ser uma tarefa difícil. Para resolver este problema decidiu-se considerar homogêneos os TIs com pelo menos 50% dos seus municípios agrupados numa região específica da grade neural

Para melhor caracterizar os TIs a partir dos Planos de Componentes foi definido um contorno para cada TI na grade neural que destaca a região que agrupa pelo menos 50% dos municípios considerados na análise do TI. Na Figura 6 tem-se a representação dos contornos que representarão cada TI considerado homogêneo, sendo que o contorno

Figura 6
Contornos dos grupos de neurônios artificiais que estão associados a seis TIs representantes daqueles com maior IHT



Fonte: Elaborado pelo autor.

Foram selecionadas somente nove variáveis para caracterizar os TIs considerados homogêneos tanto pelo IHT quanto pela análise da dispersão dos municípios de cada TI na grade neural

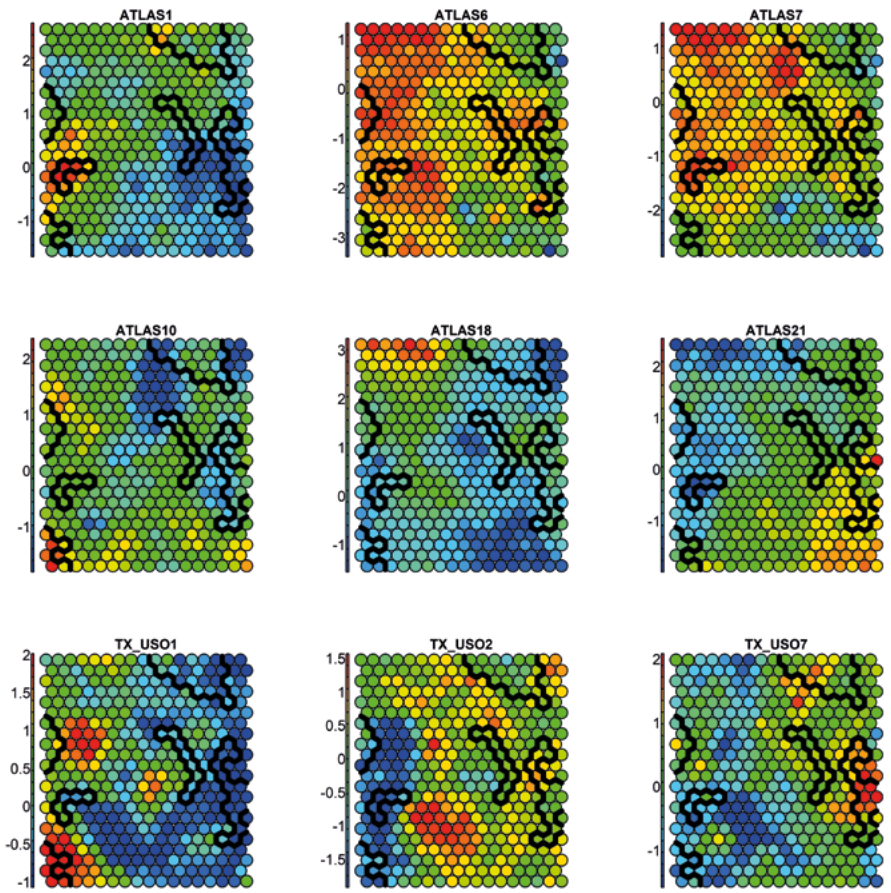
do Baixo Sul agrupa 57% de todos os municípios analisados do TI, o do TI Médio Sudoeste da Bahia 60%, o contorno do TI Extremo Sul 70%, o contorno do TI Bacia do Jacuípe 79%, o contorno do TI Piemonte Norte do Itapicuru 86% e o contorno do TI Bacia do Rio Corrente 100%.

Interpretação dos Planos de Componentes

Os Planos de Componentes para todas as variáveis encontram-se no Anexo B. No entanto, foram selecionadas somente nove variáveis para caracterizar os TIs considerados homogêneos tanto pelo IHT quanto pela análise da dispersão dos municípios de cada TI na grade neural (Figura 7). Procurou-se selecionar variáveis não correlacionadas e que ajudam a destacar as diferenças entre os seis TIs.

Observa-se pelo padrão coroplético distinto entre todos os Planos de Componentes que as variáveis escolhidas não estão correlacionadas.

Figura 7
Planos de Componentes para as variáveis ATLAS1, ATLAS6, ATLAS7, ATLAS10, ATLAS18, ATLAS21, TX_USO1, TX_USO2, TX_USO7 com os contornos dos grupos de neurônios artificiais que estão associados aos seis TIs considerados homogêneos



Fonte: Elaborado pelo autor.

Para as variáveis percentual da renda apropriada pelos 20% mais pobres (ATLAS1), percentual da população em domicílios com densidade maior que 2 (ATLAS10), IDHM Educação (ATLAS18), percentual de pobres (ATLAS21), a taxa de lavouras permanentes (TX_USO1) e a taxa de pastagens plantadas em boas condições (TX_USO7) observa-se que foi dedicada uma pequena área da grade neural especializada nos municípios com valores acima da média. O inverso ocorre para as variáveis percentual da população em domicílios com água encanada (ATLAS6) e percentual da população em domicílios com banheiro e água encanada (ATLAS7). Apenas a variável taxa de lavouras temporárias (TX_USO2) apresentou equilíbrio entre valores acima e abaixo da média.

Os seis contornos dos TIs não se cruzam e estão bem distribuídos na grade neural, o que, aliado aos diferentes padrões dos Planos de Componentes para as nove variáveis, facilita a caracterização de cada TI a partir de cada um dos Planos de Componentes. A Tabela 2 resume a interpretação dos Planos de Componentes para cada TI e variável considerados nesta caracterização.

Os seis contornos dos TIs não se cruzam e estão bem distribuídos na grade neural, o que, aliado aos diferentes padrões dos Planos de Componentes para as nove variáveis, facilita a caracterização de cada TI a partir de cada um dos Planos de Componentes

Tabela 2
Indicação de valores observados nos Planos de Componentes acima da média (↑, tonalidades de azul no Plano de Componente), abaixo da média (↓, tonalidades de vermelho no Plano de Componente) ou intermediários (↔, tonalidades de verde no Plano de Componente) para as nove variáveis analisadas e para cada um dos seis TIs considerados homogêneos

Variável (média para o estado da Bahia)	Território de Identidade (média para o TI)					
	Baixo Sul	Extremo Sul	Médio Sudoeste da Bahia	Bacia do Jacuípe	Bacia do Rio Corrente	Piemonte Norte do Itapicuru
ATLAS1 (2,6)	↔ (2,8)	↔ (2,8)	↑ (4,0)	↔ (3,1)	↓ (1,9)	↔ (2,1)
ATLAS6 (75,6)	↔ (76,5)	↑ (86,6)	↑ (81,2)	↔ (69,6)	↔ (75,4)	↑ (77,7)
ATLAS7 (67,4)	↔ (62,2)	↑ (74,1)	↑ (83,4)	↔ (61,1)	↔ (70,6)	↔ (64,6)
ATLAS10 (27,4)	↑ (37,3)	↔ (28,5)	↔ (28,4)	↔ (22,4)	↓ (20,3)	↔ (22,9)
ATLAS18 (0,318)	↓ (0,311)	↔ (0,337)	↓ (0,288)	↓ (0,280)	↔ (0,322)	↓ (0,343)
ATLAS21 (20,4)	↔ (18,2)	↓ (13,5)	↓ (11,2)	↔ (18,0)	↔ (25,2)	↔ (21,3)
TX_USO1 (0,4)	↑ (0,9)	↔ (0,4)	↔ (0,4)	↓ (0,1)	↓ (0,1)	↓ (0,2)
TX_USO2 (0,5)	↓ (0,3)	↓ (0,4)	↓ (0,2)	↔ (0,5)	↔ (0,6)	↔ (0,5)
TX_USO7 (0,4)	↓ (0,2)	↔ (0,4)	↓ (0,3)	↔ (0,5)	↑ (0,7)	↔ (0,5)

Fonte: Elaborado pelo autor.

Por exemplo, o contorno que delimita o TI Baixo Sul está localizado numa área da grade neural com valores bem acima da média para, por exemplo, a variável taxa de lavoura permanente (TX_USO1). De fato, é o único TI desses avaliados com forte presença da agricultura permanente, principalmente o cacau em Igrapiúna e a banana em Wenceslau Guimarães.

Não se questiona a pertinência da formação dos TIs, apenas se deseja apontar diferentes perspectivas sobre a homogeneidade dos TIs a partir de dados multivariados, complementando os demais estudos

A Tabela 2 também mostra que os TIs Extremo Sul e Médio Sudoeste da Bahia estão próximos na grade neural e compartilham a mesma leitura dos Planos de Componentes para quatro das variáveis. Ou seja, compartilham variáveis com valores acima da média para o percentual de população em domicílios com água encanada e banheiro (ATLAS6 e ATLAS7), que reflete o alto grau de urbanização destes TIs, e variáveis com valores abaixo da média para percentual de extremamente pobres (ATLAS21), reflexo do bom desempenho industrial (papel e celulose no Extremo Sul e calçadista no Médio Sudoeste da Bahia) e de lavouras temporárias (TX_USO2). O TI Médio Sudoeste da Bahia se destaca dos demais quanto aos altos valores para a variável percentual da renda apropriada pelos 20% mais pobres (ATLAS1), o que denota menor desigualdade social, enquanto o TI Bacia do Rio Corrente se destaca pelo inverso.

O TI Bacia do Rio Corrente se destaca dos demais pelo baixo percentual da renda apropriada pelos 20% mais pobres (ATLAS1), baixo percentual da população em domicílios com densidade maior que dois (ATLAS10) o que reflete a baixa urbanização do TI, e baixa presença de lavouras permanentes (TX_USO1) pois predomina as culturas temporárias de soja, milho e cana-de-açúcar. Este mesmo TI se destaca pela forte presença de pastagens plantadas em boas condições (TX_USO7).

Os Planos de Componentes mostraram que o TI Bacia do Jacuípe apresenta valores próximos a média da Bahia para quase todas as variáveis, exceto para os valores abaixo da média para o IDHM-Educação (ATLAS18), que reflete as altas taxas de analfabetismo e pobreza, e lavouras permanentes (TX_USO1).

O TI Piemonte Norte do Itapicuru se destaca no percentual elevado da população em domicílios com água encanada (ATLAS6), apesar da baixa urbanização, baixo índice de educação IDHM-Educação (ATLAS18) e pouca presença de lavouras permanentes (TX_USO1), pois na região semiárida destaca-se a produção de caprinos e ovinos (SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAIS DA BAHIA, 2018).

Esta análise não exaustiva das características dos TIs considerados homogêneos, ou pelo menos mais homogêneos que os demais, mostrou que há coerência entre os valores calculados para o Índice de Homogeneidade Territorial e a interpretação da projeção dos municípios de cada TI na grade neural e posterior interpretação rápida pelos Planos de Componentes. Importante destacar que não se questiona a pertinência da formação dos TIs, apenas se deseja apontar diferentes perspectivas sobre a homogeneidade dos TIs a partir de dados multivariados, complementando os demais estudos (SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAS DA BAHIA, 2015, 2016, 2018; FIGUEIRA; FIGUEIRA, 2017; SANTOS; SILVA; PEREIRA, 2011; MONTEIRO; SERPA, 2011).

CONCLUSÕES

A propriedade de ordenação topológica da RNA SOM permitiu uma rápida caracterização dos Territórios de Identidade a partir dos Planos de Componentes, identificando variáveis correlacionadas, a intensidade da variação e a dispersão das observações. A aplicação desta técnica pode preceder uma análise de componentes principais (para dados quantitativos) ou uma análise de correspondência múltipla (para dados categóricos), auxiliando no processo de elaboração de hipóteses sobre os dados. Assim como, se aproveitar destes métodos estatísticos para seleção das variáveis que serão avaliadas pela RNA. Adicionalmente, é possível aplicar esta técnica de visualização para análise de tendências para séries temporais.

A análise visual dos dados dos TIs complementou a avaliação da homogeneidade a partir do Índice de Homogeneidade Territorial, confirmando a heterogeneidade dos TIs com baixo IHT como Chapada Diamantina, Litoral Norte e Agreste Baiano e Portal do Sertão, e mostrando que há dissimilaridades mesmo em TIs com altos valores para o IHT, como é o caso dos TIs Baixo Sul, Bacia do Jacuípe, Extremo Sul, Bacia do Rio Corrente, Piemonte Norte do Itapicuru e Médio Sudoeste da Bahia.

A avaliação da homogeneidade/heterogeneidade territorial é de fundamental importância para a elaboração, monitoramento e avaliação de políticas públicas territoriais. E esta avaliação deve estar amparada em estudos interdisciplinares que articulem diferentes métodos de análise e que levem em consideração os diferentes pontos de vista possíveis sobre a questão da identidade territorial.

A avaliação da homogeneidade/heterogeneidade territorial é de fundamental importância para a elaboração, monitoramento e avaliação de políticas públicas territoriais

REFERÊNCIAS

BAHIA. Decreto nº 12.354 de 25 de agosto de 2010. Institui o Programa Territórios de Identidade e dá outras providências. *Diário Oficial [do] Estado da Bahia*, Salvador, 26 ago. 2010.

BLATT, N.; GONDIM, P. S. C. Territórios de identidade no estado da Bahia: uma análise da regionalização implantada pela estrutura governamental na perspectiva do desenvolvimento local e regional. *Colóquio Baiano espaços, Tempo, Espaços e Representações*: Abordagens Históricas e Geográficas, Vitória da Conquista, v. 1, n. 1, 2013.

BORGES, S. S.; SERPA, A. O papel dos agentes públicos e da sociedade civil na implementação de políticas de desenvolvimento territorial no estado da Bahia: uma análise preliminar. *Revista Geografares*, Vitória, n. 11, p. 31-59, jun. 2012.

CENSO AGROPECUÁRIO 2006: segunda apuração. Rio de Janeiro: IBGE, 2006. Disponível em: https://sidra.ibge.gov.br/pesquisa/censo-agropecuario/censo-agropecuario-2006/segunda-apuracao. Acesso em: 25 ago. 2015.

CENSO DEMOGRÁFICO 2010: resultados do universo: características da população e dos domicílios. Rio de Janeiro: IBGE, 2012. Disponível em: https://sidra.ibge.gov.br/pesquisa/censo-demografico/demografico-2010/universo-caracteristicas-da-populacao-e-dos-domicilios. Acesso em: 25 ago. 2015.

FIGUEIRA, W. A; FIGUEIRA, E. A. As implicações dos programas de transferência de renda no IDH-M do Território de Identidade do Sudoeste Baiano. *Reflexões Econômicas*, Ilhéus, v.1, n.3, p. 93-111, mar. 2017.

FIGUEIREDO, A. K. S. *et al*. Análise espacial do desenvolvimento e das desigualdades no Território Sudoeste Baiano. *Desenvolvimento Em Questão*, Ijuí, v. 16, n. 44, p. 69-104, jul./set. 2018.

INSTITUTO DE PESQUISA ECONÔMICA APLICADA. *Programa Bolsa Família* (PBF). Brasília: IPEA. 2007. Disponível em: http://www.ipeadata.gov.br. Acesso em: 14 ago. 2015.

KOHONEN, T. Essentials of the self-organizing map. *Neural Networks, London*, v. 37, p. 52–65, 2013.Disponível em: https://doi.org/doi:10.1016/j.neunet.2012.09.018. Acesso em:

KOHONEN, T. *Self-organizing maps*. 3. ed. Berlin: Springer, 2001.

MONTEIRO, J.; SERPA, A. Políticas de desenvolvimento territorial e cultural no território de identidade de Vitória da Conquista: uma análise geográfica da lógica de localização de projetos e recursos. *Ateliê Geográfico*, Goiânia, v. 5, n. 3. p. 150-171, dez. 2011.

PROGRAMA DAS NAÇÕES UNIDAS PARA O DESENVOLVIMENTO. *Atlas do desenvolvimento humano*. Brasília: PNAD, 2010. Disponível em: http://www.atlasbrasil.org.br/. Acesso em: 10 out. 2015.

SANTOS, J. P. C.; SILVA, K. M. das G. C.; PEREIRA, S. B. M. Tipologia dos municípios baianos com base em análise multivariada. Salvador: SEI, 2011. (Textos para discussão, n. n. 2). Disponível em: http://www.sei.ba.gov.br/images/publicacoes/download/textos_discussao/texto_discussao_02.pdf. Acesso em: 3 jan. 2019.

SABOURIN, E. Evolução da política federal de desenvolvimento territorial no Brasil. *Novos Cadernos NAEA*, Belém, v. 18, n. 1. p. 123-143, jan./jun. 2015.

SAQUET, M. A. *Abordagens de concepções de território*. 2. ed. São Paulo: Expressão Popular, 2010.

SILVA, M. A. S. Dados multivariados utilizados para o cálculo do IHT dos Territórios de Identidade. *GeoInfo* [Data Set], 2019. Primeira versão. Disponível em: http://inde.geoinfo.cnpm.embrapa.br/geonetwork_inde/srv/por/catalog.search#/metadata/6d8c345a-f974-11e9-befa-0200753f7c0c. Acesso em: 19 jul. 2020.

SILVA, M. A. S.; SOUZA, R. A. Avaliação da homogeneidade dos Territórios de Identidade a partir de técnicas geocomputacionais. *Revista Brasileira de Desenvolvimento Regional*, Blumenau, v. 6, n. 3, p. 111-146, 2018.

SILVA, M. A. S. *et al*.. Using self-organizing maps for rural territorial typology. *In*: PRADO, H. A. do; BARRETO LUIZ, Alfredo Jose; HOMERO FILHO, Chaib. (org.). *Computational methods for agricultural research*: advances and applications. Hershey: Information Science Reference, 2011. p. 107-126. Disponível em: https://doi.org/10.4018/978-1-61692-871-1.ch007"871-1.ch007. Acesso em: 19 jul. 2020.

SILVA, M. A. S. *et al*.. *TerraSOM*: Sistema para Análise de Dados Geoespaciais Agregados por Área Baseado na Rede Neural do Tipo Mapa Auto-Organizável de Kohonen. Aracaju: Embrapa Tabuleiros Costeiros, 2015. (Boletim de pesquisa e desenvolvimento, n. 65).

SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAIS DA BAHIA. *Perfil dos Territórios de Identidade da Bahia*. Salvador: SEI, 2015. v. 1.

SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAIS DA BAHIA. *Perfil dos Territórios de Identidade da Bahia*. Salvador: SEI, 2016. v. 2.

SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAIS DA BAHIA. *Perfil dos Territórios de Identidade da Bahia*. Salvador: SEI, 2018. v. 3.

VELLOSO, T. R. *Uma nova institucionalidade do desenvolvimento rural*: a trajetória dos territórios rurais no Estado da Bahia. 2013. Tese (Geografia) - Universidade Federal de Sergipe, Aracaju, 2013.

WEHRENS, R.; BUYDENS, L.M.C. Self- and super-organizing maps in R: the kohonen Package. *Journal of Statistical Software*, Innsbruck, v. 21, n. 5, 2007. Disponível em: https://doi.org/ 10.18637/jss.v021.i05. Acesso em: 10 jan. 2014.

WEHRENS, R.; KRUISSELBRINK, J. Flexible self-organizing maps in kohonen 3.0. *Journal of Statistical Software*, Innsbruck, v. 87, n. 7, 2018. Disponívem em: https://doi.org/10.18637/jss.v087.i07. Acesso em: 20 jan. 2019.

ANEXO A

(Continua)

Tabela A1
Lista das 45 variáveis socioeconômicas descritivas dos aspectos socioeconômico, condição do produtor rural, uso da terra e efetivo de animais

Aspecto	Ord	Descrição da variável	Sigla
Socioeconômico	1	Percentual da renda apropriada pelos 20% mais pobres	ATLAS1
	2	Razão 10% mais ricos / 40% mais pobres	ATLAS2
	3	Índice de Theil – L	ATLAS3
	4	Índice de Gini	ATLAS4
	5	Percentual dos ocupados com fundamental completo – 18 anos ou mais	ATLAS5
	6	Percentual da população em domicílios com água encanada	ATLAS6
	7	Percentual da população em domicílios com banheiro e água encanada	ATLAS7
	8	Percentual da população em domicílios com coleta de lixo	ATLAS8
	9	Percentual da população em domicílios com energia elétrica	ATLAS9
	10	Percentual da população em domicílios com densidade > 2	ATLAS10
	11	População rural / População total	TX_POPRURA
	12	População urbana / População total	TX_POPURBA
	13	Mortalidade infantil	ATLAS14
	14	IDHM Renda	ATLAS15
	15	IDHM Longevidade	ATLAS16
	16	Subíndice de frequência escolar – IDHM Educação	ATLAS17
	17	Subíndice de escolaridade – IDHM Educação	ATLAS18
	18	Taxa de fecundidade total	ATLAS19
	19	Taxa de analfabetismo – 18 anos ou mais	ATLAS20
	20	Percentual de extremamente pobres	ATLAS21
	21	Percentual de pobres	ATLAS22
	22	Programa Bolsa Família (PBF) – valor total dos benefícios em dezembro (2007) / número de benefícios em dezembro (2007)	TX_BF_2007
Condição do produtor rural	23	Proprietário / Total de número de estabelecimentos	TX_PROPRIE
	24	Arrendatário/ Total de número de estabelecimentos	TX_ARRENDA
	25	Parceiro / Total de número de estabelecimentos	TX_PARCEIR
	26	Ocupante / Total de número de estabelecimentos	TX_OCUPANT
	27	Produtor sem área / Total de número de estabelecimentos	TX_PRODUTO
Uso da terra	28	Quantidade produzida de leite de vaca no ano (Mil litros) / 60% das Vacas ordenhadas no ano (Cabeças)	TX_LEITEVA
	29	Lavouras – permanentes / Total de número de estabelecimentos	TX_USO1
	30	Lavouras – temporárias / Total de número de estabelecimentos	TX_USO2
	31	Lavouras – área plantada com forrageiras para corte / Total de número de estabelecimentos	TX_USO3
	32	Pastagens – naturais / Total de número de estabelecimentos	TX_USO5
	33	Pastagens – plantadas degradadas / Total de número de estabelecimentos	TX_USO6
	34	Pastagens – plantadas em boas condições / Total de número de estabelecimentos	TX_USO7
	35	Matas e/ou florestas – naturais destinadas à preservação permanente ou reserva legal / Total de número de estabelecimentos	TX_USO8
	36	Matas e/ou florestas – naturais (exclusive área de preservação permanente e as em sistemas agroflorestais) / Total de número de estabelecimentos	TX_USO9
	37	Matas e/ou florestas – florestas plantadas com essências florestais / Total de número de estabelecimentos	TX_USO10

(Conclusão)

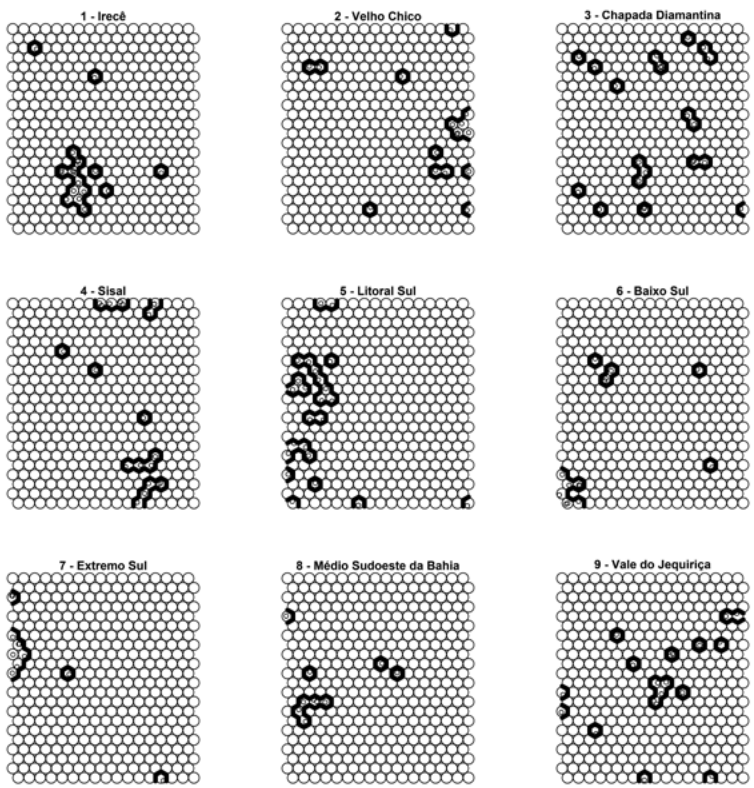
Tabela A1
Lista das 45 variáveis socioeconômicas descritivas dos aspectos socioeconômico, condição do produtor rural, uso da terra e efetivo de animais

Aspecto	Ord	Descrição da variável	Sigla
	38	Sistemas agroflorestais – área cultivada com espécies florestais também usadas para lavouras e pastoreio por animais / Total de número de estabelecimentos	TX_USO11
	39	Tanques, lagos, açudes e/ou área de águas públicas para exploração da aquicultura / Total do número de estabelecimentos	TX_USO12
	40	Construções, benfeitorias ou caminhos / Total de número de estabelecimentos	TX_USO13
	41	Terras degradadas (erodidas, desertificadas, salinizadas etc.) / Total de número de estabelecimentos	TX_USO14
	42	Terras inaproveitáveis para agricultura ou pecuária (pântanos, areais, pedreiras etc.) / Total de número de estabelecimentos	TX_USO15
Efetivo de animais	43	Número de cabeças de bovinos (Cabeças) / Número de estabelecimentos agropecuários com efetivo de bovinos em 31/12 (Unidades)	TX_BOVINO
	44	Número de cabeças de caprinos (Cabeças) / Número de estabelecimentos agropecuários com caprinos (Unidades)	TX_CAPRINO
	45	Número de cabeças de ovinos (Cabeças) / Número de estabelecimentos agropecuários com ovinos (Unidades)	TX_OVINO

Fonte: Adaptada de Silva e Souza (2018).

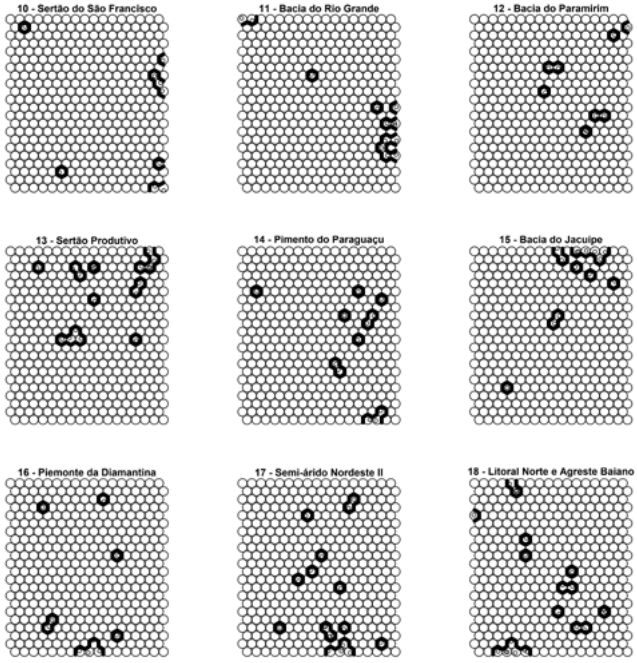
ANEXO B

Figura B1
Distribuição dos TIs de 1 a 9 na grade neural artificial após o processo de aprendizagem de máquina



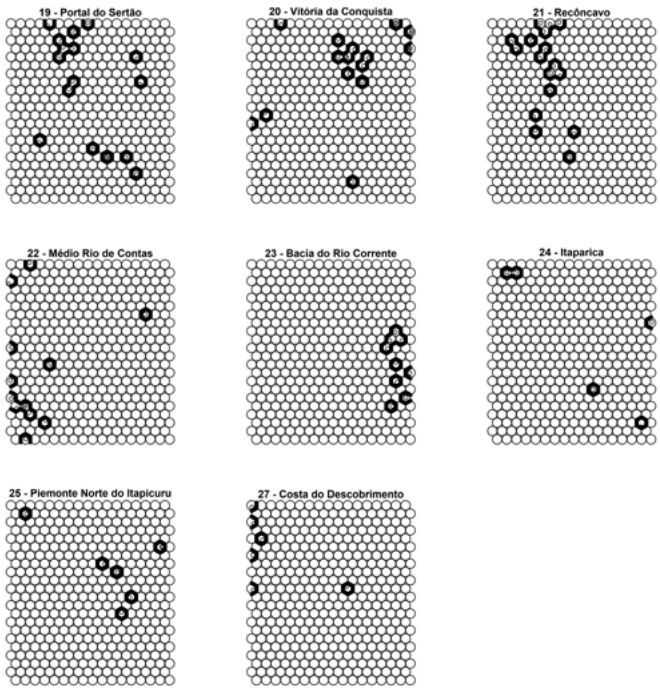
Fonte: Elaborado pelo autor.

Figura B2
Distribuição dos TIs de 10 a 18 na grade neural artificial após o processo de aprendizagem de máquina



Fonte: Elaborado pelo autor.

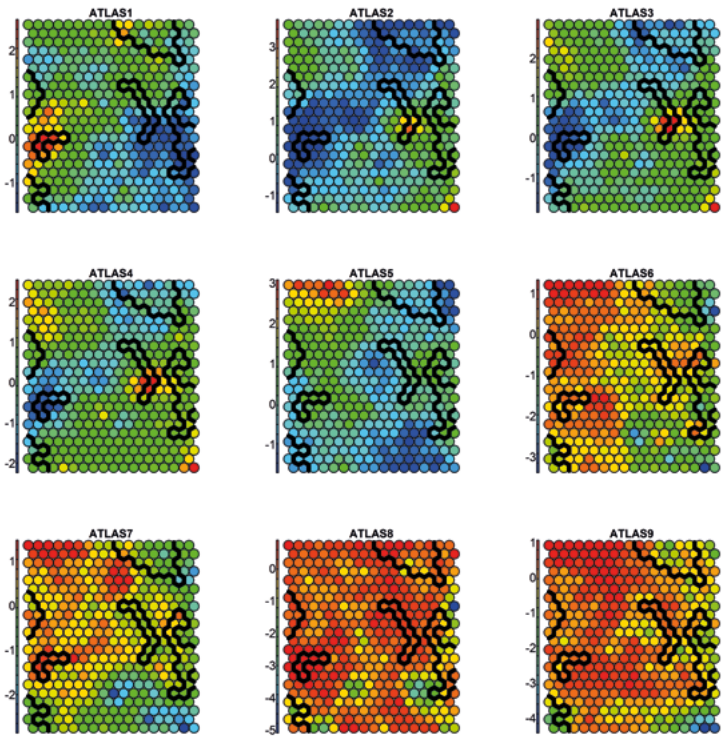
Figura B3
Distribuição dos TIs de 19 a 27 na grade neural artificial após o processo de aprendizagem de máquina



Fonte: Elaborado pelo autor.

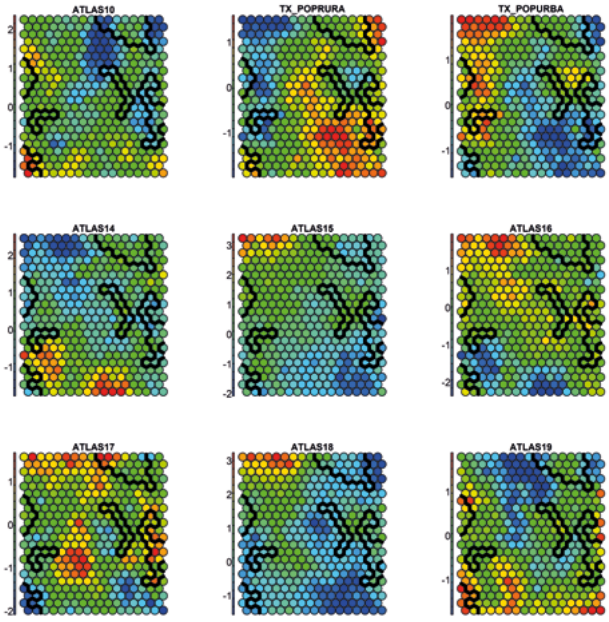
ANEXO C

Figura C1
Planos de Componentes das variáveis de estudo (1 a 9) considerando os valores dos pesos de cada neurônio artificial representados por cores divergentes (azul representa valores abaixo da média, e vermelho acima da média). Também estão representados na grade neural artificial seis dos TIs com maior homogeneidade territorial segundo o índice IHT



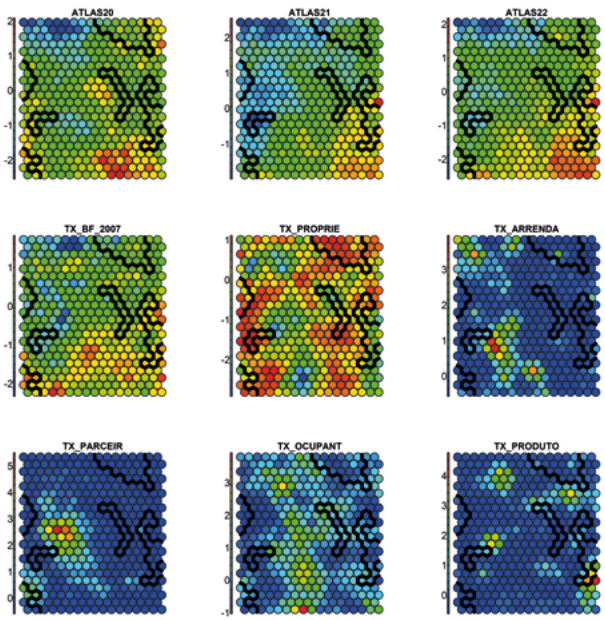
Fonte: Elaborado pelo autor.

Figura C2
Planos de Componentes das variáveis de estudo (10 a 18) considerando os valores dos pesos de cada neurônio artificial representados por cores divergentes (azul representa valores abaixo da média, e vermelho acima da média). Também estão representados na grade neural artificial seis dos TIs com maior homogeneidade territorial segundo o índice IHT



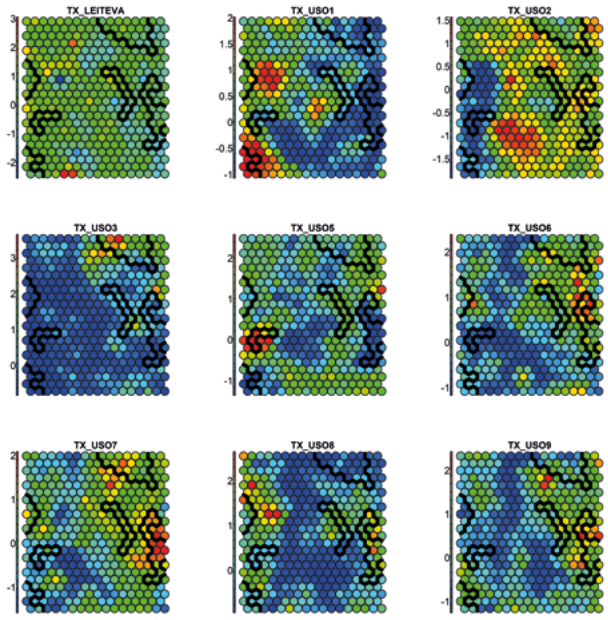
Fonte: Elaborado pelo autor.

Figura C3
Planos de Componentes das variáveis de estudo (19 a 27) considerando os valores dos pesos de cada neurônio artificial representados por cores divergentes (azul representa valores abaixo da média, e vermelho acima da média). Também estão representados na grade neural artificial seis dos TIs com maior homogeneidade territorial segundo o índice IHT



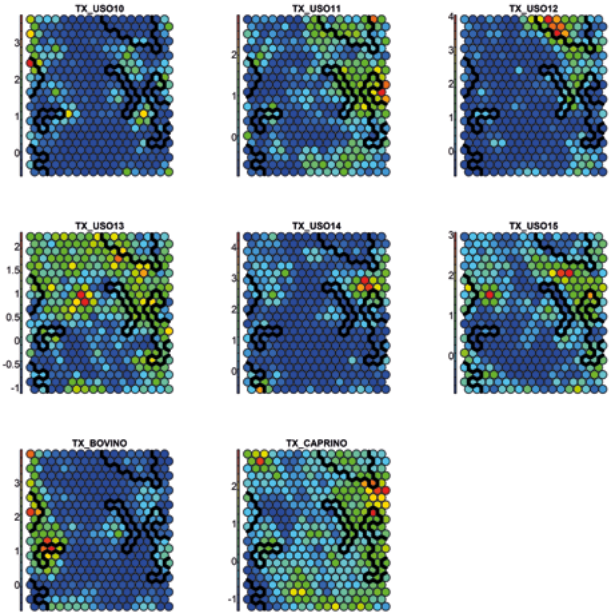
Fonte: Elaborado pelo autor.

Figura C4
Planos de Componentes das variáveis de estudo (28 a 36) considerando os valores dos pesos de cada neurônio artificial representados por cores divergentes (azul representa valores abaixo da média, e vermelho acima da média). Também estão representados na grade neural artificial seis dos TIs com maior homogeneidade territorial segundo o índice IHT



Fonte: Elaborado pelo autor.

Figura C5
Planos de Componentes das variáveis de estudo (37 a 45) considerando os valores dos pesos de cada neurônio artificial representados por cores divergentes (azul representa valores abaixo da média, e vermelho acima da média). Também estão representados na grade neural artificial seis dos TIs com maior homogeneidade territorial segundo o índice IHT



Fonte: Elaborado pelo autor.

NORMAS PARA PUBLICAÇÃO

A revista *Bahia Análise & Dados*, editada pela Superintendência de Estudos Econômicos e Sociais da Bahia (SEI), órgão vinculado à Secretaria do Planejamento do Estado da Bahia (Seplan), aceita colaborações originais, em português, inglês e espanhol, de artigos sobre os temas definidos nos editais publicados no site da SEI, bem como resenhas de livros inéditos que se enquadrem no tema correspondente.

Os artigos e resenhas são submetidos à apreciação do conselho editorial, instância que decide sobre a publicação.

A editoria da SEI e a coordenação editorial da edição reservam-se o direito de sugerir ou modificar títulos, formatar tabelas e ilustrações, dentre outras intervenções, a fim de atender ao padrão editorial e ortográfico adotado pela instituição, constante no Manual de Redação e Estilo da SEI, disponível no site www.sei.ba.gov.br, menu "Publicações".

Os artigos ou resenhas que não estiverem de acordo com as normas não serão apreciados.

O autor terá direito a um exemplar do periódico em que seu artigo for publicado.

PADRÃO PARA ENVIO DE ARTIGOS OU RESENHAS

- Artigos e resenhas devem ser enviados, preferencialmente, através do site da revista, opção "Submissão", ou pelo e-mail definido no edital, para a coordenação editorial daquele número.
- Devem ser apresentados em editor de texto de maior difusão (Word), formatados com entrelinhas de 1,5, margem esquerda de 3 cm, direita e inferior de 2 cm, superior de 3 cm, fonte Times New Roman, tamanho 12.
- Devem ser assinados, preferencialmente, por, no máximo, três autores.
- É permitido apenas um artigo por autor, exceto no caso de participação como coautor.
- O autor deve incluir, em nota de rodapé, sua identificação, com nome completo, titulação acadêmica, nome da(s) instituição(ões) a que está vinculado, e-mail, telefone e endereço para correspondência.
- Os artigos devem conter, no mínimo, 15 páginas e, no máximo, 25, e as resenhas, no máximo, três páginas.
- Devem vir acompanhados de resumo e *abstract* contendo de 100 a 250 palavras, ressaltando o objetivo, a metodologia, os principais resultados e a conclusão. Palavras-chave e *keywords* devem figurar abaixo, separadas por ponto e finalizadas também com ponto.
- Apresentar padronização de título, de forma a ficar claro o que é título e subtítulo. O título deve se constituir de palavra, expressão ou frase que designe o assunto ou conteúdo do texto. O subtítulo, apresentado em seguida ao título e dele separado por dois pontos, visa esclarecê-lo ou complementá-lo.
- As tabelas e demais ilustrações (desenhos, esquemas, figuras, fluxogramas, fotos, gráficos, mapas etc.) devem estar numeradas consecutivamente, com algarismos arábicos, na ordem em que forem citadas no texto, com os títulos, legendas e fontes completas, e localizadas o mais próximo possível do trecho a que se referem.
- Tabelas e gráficos devem ser enviados em programa de planilhas de maior difusão (Excel). Fotografias e ilustrações escaneadas devem apresentar resolução de 300 dpi (CMYK), com cor real e salvas na extensão TIFF.
- As citações de até três linhas devem estar entre aspas, na sequência do texto. As citações com mais de três linhas devem constar em parágrafo próprio, com recuo da margem de 4 cm, fonte 10, espaço simples, sem aspas e identificadas pelo sistema autor-data (NBR 10520 da ABNT).
- Quando da inclusão de depoimentos dos sujeitos, apresentá-los em parágrafo distinto do texto, entre aspas, com letra e espaçamento igual ao do texto e recuo esquerdo, de todas as linhas, igual ao do parágrafo.
- As notas de rodapé devem ser explicativas ou complementares, curtas, numeradas em ordem sequencial, no corpo do texto e na mesma página em que forem citadas.
- As referências devem ser completas e precisas, segundo as Normas Brasileiras para Referências Bibliográficas – NBR 6023 da ABNT.

Todos os números da *Bahia Análise & Dados* podem ser visualizados no site da SEI (www.sei.ba.gov.br) no menu "Publicações".

SEI 25 ANOS DE SEI
65 ANOS DE HISTÓRIA
SUPERINTENDÊNCIA DE ESTUDOS ECONÔMICOS E SOCIAIS DA BAHIA

GOVERNO DO ESTADO
SECRETARIA DO PLANEJAMENTO